C02029: Doctor of Philosophy
32903 PhD Thesis: Analytics
December 2021

Customer Behavior Analytics and Visualization

Md Rafiqul Islam

School of Computer Science
Faculty of Engineering and Information Technology
University of Technology Sydney
NSW - 2007, Australia

Customer Behavior Analytics and Visualization

by Md Rafiqul Islam

Thesis submitted in fulfilment of the requirements for the degree of

Doctor of Philosophy
in
Analytics

Under the supervision of Professor Guandong Xu and Dr. Xianzhi Wang

School of Computer Science
Faculty of Engineering and Information Technology
University of Technology Sydney
NSW - 2007, Australia

December 2021

CERTIFICATE OF ORIGINAL AUTHORSHIP

I, **Md Rafiqul Islam** declare that this thesis, is submitted in fulfilment of the requirements for the award of **Doctor of Philosophy in Analytics**, in the **School of Computer Science**, **Faculty of Engineering and Information Technology at the University of Technology Sydney, Australia**.

This thesis is wholly my own work unless otherwise referenced or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

This document has not been submitted for qualifications at any other academic institution.

This research is supported by the Australian Government Research Training Program.

Production Note:
SIGNATURE: Signature removed prior to publication.

[Md Rafiqul Islam]

DATE: 14th December, 2021

PLACE: Sydney, Australia

DEDICATION

To my beloved parents and teachers

Acknowledgments

First of all, I owe my profound gratitude to my supervisor, Prof. Guandong Xu for providing me with all the resources, scholarly advice, and magnanimous technical, and financial support throughout my Ph.D. study. As a mentor, Prof. Guandong Xu showed me the quality of being devoted to my work through his contagious passion for research by being a role model.

I would also like to convey my appreciation and gratitude to my co-supervisor Dr. Xianzhi Wang for his insightful comments and encouragement. Besides my supervisors, my sincere appreciation goes to external collaborators Dr. Imran Razzak and Dr. Shaowu Liu who provided me essential support during my research work. Without their precious support, it would be difficult to achieve this goal.

I sincerely acknowledge my dear teacher Prof. Dr. Abu Raihan M. Kamal, Dr. Ashad Kabir, Dr. Md. Samiullah, Dr. Anwar Ulhaq, and friends Dr. Munshi Muhammad Abdul Kader Jilani, Dr. Mohammad Muntaseer Mahfuz, Dr. Md Shamsur Rahim, Dr. Shakil Ahmed Khan, Dr. Pejush Chandra Sarker, Shanjita Akter, Md. Ashraful Haque, and lab mates who have been extremely supportive during the entire course of my Ph.D. study.

I would like to acknowledge full financial support through the International Research Scholarship (IRS) and Faculty of Engineering and Information Technology (FEIT) Scholarship throughout my Ph.D. study.

Last but not the least, my heartfelt appreciation goes to all my family members, especially my father Md Abdul Kader, mother Most Tahmina Begum, and my uncle Md Fazlul Haque who have been extremely supportive during the entire course of my Ph.D. I would also extend my gratitude to my wife Aireen Rahman. Without her cooperation and support, it would not have been possible to accomplish this goal.

List of Publications

List of Journal Articles (Published/Accepted)

1. **Islam, Md Rafiqul**, Imran Razzak, Xianzhi Wang, Peter Tilocca, and Guandong Xu. "Natural language interactions enhanced by data visualization to explore insurance claims and manage risk." Annals of Operations Research (2022): 1-19. (**Refer to Chapter 1 & 5**)

2. **Islam, Md Rafiqul**, Shanjita Akter, Md Rakybuzzaman Ratan, Linta Islam, Imran Razzak, and Guandong Xu. “Strategies for evaluating visual interactive system: a systematic review and new perspectives." Journal of Visualization, (2022). (**Refer to Chapter 1 & 6**)

3. **Islam, Md Rafiqul**, Shaowu Liu, Rhys Biddle, Imran Razzak, Xianzhi Wang, Peter Tilocca, and Guandong Xu. “Discovering dynamic adverse behavior of policyholders in the life insurance industry." Technological Forecasting and Social Change 163 (2021): 120486. (**Refer to Chapter 1 & 3**)

4. **Islam, Md Rafiqul**, Shanjita Akter, Md Rakybuzzaman Ratan, Abu Raihan M. Kamal, and Guandong Xu. “Deep visual analytics (dva): applications, challenges and future directions." Hum-Centric Intell Syst 1, no. 1-2 (2021): 3-17. (**Refer to Chapter 1 & 6**)

5. Mosiur Rahman, **Md Rafiqul Islam**, Sharmin Akter, Shanjita Akter, Linta Islam, and Guandong Xu. “DiaVis: exploration and analysis of diabetes through visual interactive system." Hum-Centric Intell Syst (2021). (**Refer to Chapter 4**)

6. Sharif, Omar, **Md Rafiqul Islam**, Md Zobaer Hasan, Muhammad Ashad Kabir, Md Emran Hasan, Salman A. AlQahtani, and Guandong Xu. “Analyzing the impact of demographic variables on spreading and forecasting COVID-19." Journal of Healthcare Informatics Research (2021): 1-19. (**Refer to Chapter 2**)

7. **Islam, Md Rafiqul**, Shaowu Liu, Xianzhi Wang, and Guandong Xu. “Deep learning for misinformation detection on online social networks: a survey and new perspectives." Social Network Analysis and Mining 10, no. 1 (2020): 1-20. (**Refer to Chapter 2**)

List of Conference Articles (Published)

1. **Islam, Md Rafiqul**, Imran Razzak, Xianzhi Wang, Peter Tilocca, and Guandong Xu. “UCBVis: understanding customer behavior sequences with visual interactive system." In 2021 International Joint Conference on Neural Networks (IJCNN), pp. 1-8. IEEE, 2021. (**Refer to Chapter 1 & 6**)

2. **Islam, Md Rafiqul**, Jiaming Zhang, Md. Hamjajul Ashmafee, Imran Razzak, Jianlong Zhou, Xianzhi Wang, and Guandong Xu. “ExVis: explainable visual decision support system for risk management." In 2021 8th International Conference on Behavioural and Social Computing (BESC), IEEE, 2021. (**Refer to Chapter 4**)

3. Zerafa, Joshua, **Md Rafiqul Islam**, Ashad Kabir, and Guandong Xu. “ExTraVis: exploration of traffic incidents using visual interactive system." In 25th International Conference Information Visualisation (IV 2021). IEEE, Institute of Electrical and Electronics Engineers, 2021. (**Refer to Chapter 2**)

4. **Islam, Md Rafiqul**, Shaowu Liu, Imran Razzak, Muhammad Ashad Kabir, Xianzhi Wang, Peter Tilocca, and Guandong Xu. “MHIVis: visual analytics for exploring mental illness of policyholders in life insurance industry." In 2020 7th International Conference on Behavioural and Social Computing (BESC), pp. 1-4. IEEE, 2020. (**Refer to Chapter 4**)

Table of Contents

Certificate of Original Authorship **ii**

Acknowledgments **vii**

List of Publications **ix**

Table of Contents **xii**

List of Figures **xiii**

List of Tables **xv**

Abstract **xix**

1 Introduction **1**

1.1 Background and Motivation 1

1.1.1 Analytics for Decision Making 1

1.1.2 Customer Behavior Discovery 2

1.1.3 Customer Behavior Visualization 3

1.1.4 Summary 3

1.2 Research Objectives 4

1.3 Research Problems 5

1.4 Research Contributions 7

1.5 Thesis Organization 8

2 Related Work **11**

2.1 Literature Search Methodology 11

2.2 Understanding Customer Behavior 12

2.2.1 Categorization of Customer Behavior 13

2.2.2 Current Customer Behavior Issues 16

2.2.3 Summary 18
2.3 Visualizing Customer Behavior 19
2.4 Customer Behavior Analysis using Data Mining Techniques 19
2.5 Summary 21

3 Pattern Mining for Discovering Customer Adverse Behavior 23
3.1 Background and Motivation 23
3.2 Preliminary on Adverse Behavior Selection 26
3.3 Methodology 28
3.3.1 Data Collection and Processing 28
3.3.2 Proposed Model 29
3.4 Experiment 33
3.4.1 Data Description 33
3.4.2 Interesting Pattern List Construction 34
3.5 Result and Discussion 35
3.5.1 Result Analysis 35
3.5.2 Discussion and Implications 41
3.6 Summary 43

4 Visual Analytics for Exploring Customer Behavior 45
4.1 Background and Motivation 45
4.2 Preliminary on Exploring Customer Behavior 50
4.2.1 Explainable Visualization System for Risk Management 50
4.2.2 Exploring Mental Illness of Customer with Visual Interactive System 52
4.2.3 Exploring Diabetes through Visual Interactive System 53
4.2.4 Summary 55
4.3 Methodology 56
4.3.1 Methodology of *ExVis* 56
4.3.2 Methodology of *MHIVis* 60
4.3.3 Methodology of *DiaVis* 63
4.4 Experimental Analysis 67
4.4.1 Experimental Analysis of *ExVis* 67
4.4.2 Experimental Analysis of *MHIVis* 67
4.4.3 Experimental Analysis of *DiaVis* 69
4.5 Result and Discussion 72
4.5.1 Result Analysis 73

4.5.2 User Study 74
4.5.3 Discussion 74
4.6 Summary 79

5 NLI-driven-DV to Explore Customer Claim Behavior and Manage Risk 81
5.1 Background and Motivation 81
5.2 Preliminary on Exploring Customer Claim Behavior and Manage Risk . . 84
5.2.1 Insurance Claims and Risk Management 85
5.2.2 Visualization for Claim and Risk Management 85
5.2.3 Natural Language Interfaces with Data Visualization 86
5.3 Methodology 87
5.3.1 Data Description 87
5.3.2 Data Pre-processing 88
5.3.3 Domain Characterization and Design Consideration 88
5.3.4 Visual Analytics Solution 90
5.4 Experimental Analysis 91
5.4.1 Data Interpretation 92
5.4.2 Query Analyzer 93
5.4.3 Visualization Generation 93
5.4.4 Implementation 94
5.5 Result and Discussion 95
5.5.1 Result Analysis 95
5.5.2 User Study 96
5.5.3 Discussion 99
5.6 Summary 101

6 Deep Visual Analytics for Understanding Customer Behavior 103
6.1 Background and Motivation 103
6.2 Preliminary on Deep Visual Analytics for Understanding Customer Behavior107
6.2.1 Exploring Customer Behavior with Visual Interactive System . . . 107
6.2.2 Deep Visual Analytics 110
6.2.3 Strategies for Evaluating Visualization System 113
6.3 Methodology 114
6.3.1 Methodology of *UCBVis* 114
6.3.2 Methodology of *Multi-DLMPVis* 118
6.4 Experimental Analysis 120

6.4.1 Experimental Analysis of *UCBVis* 120
6.4.2 Experimental Analysis of *Multi-DLMPVis* 121
6.5 Discussion 123
6.5.1 Evaluation 123
6.5.2 Challenges 123
6.5.3 Future Directions 126
6.6 Summary 127

7 Conclusion and Future Work **129**
7.1 Contributions 129
7.2 Future work 131

Bibliography **133**

LIST OF FIGURES

FIGURE NAME **Page**

2.1 Proposed review methodology for sample collection and analysis. 12

3.1 The structure of the *ARLAS* framework. 29

3.2 Identifying different rules with different support threshold. 36

3.3 Number of adverse policyholders per state. 37

3.4 Number of policyholders per state. 37

3.5 Number of adverse policyholder by occupation and gender. 39

3.6 Number of policyholders by age group. 40

3.7 Number of adverse policyholders by age group. 40

4.1 *ExVis*: exploring multiple views to explain visual decision support system. . 46

4.2 *MHIVis*: a visual interactive system for exploring mental health illness. . . . 47

4.3 *DiaVis*: a visual interactive system for exploring diabetes disease. 48

4.4 The framework of *ExVis* system. 57

4.5 The structure of bayesian network model. 58

4.6 The framework of *MHIVis* system. 61

4.7 The methodological framework of *DiaVis*. 64

4.8 The conceptual illustration of the study. 66

4.9 The mapping view of respondents. 71

4.10 The regional view of respondents. 71

4.11 The correlation between SBP and DBP. 72

4.12 The correlation between Height and Weight. 72

4.13 Similarity matrix of the attributes of *DiaVis*. 78

5.1 Summary of *InsCRMVis* interface components (Insurance claim and risk management with visual analytics). (a) query searching, (b) speech : allow users to freely express query to get visual insights, and (c) mouse/Touch/Pen: can be supported with visualization. 83
5.2 The data sources of the policyholders claims. 87
5.3 Proposed architecture of NLI based visualization. 91
5.4 System overview: key questions of the risk visualization framework. 92
5.5 Framework: overview of the interface functionalities such as input data, query analyzer, and visualization generation. 92
5.6 An illustration of a query analyzer while interpreting NL queries. 93
5.7 Sample questions with answers generated by the system. The answer in Q1(a), and Q1(b) is for query searching, Q2 is for speech and Q3(a) and Q3(b) is for touch/pen/mouse. 96
5.8 Post-study ranking for output visualization and opinion position. 99

6.1 Summary of *UCBVis* interface components. (a) behavior exploration workflow: allow user to identify potential behavior of customer, (b) frequent pattern view: allow user to choose sequential patterns for the focus, (c) attribute pair view: allow user to search attribute names with behavioral patterns, and (d) raw sequences view: can be supported with visualization. 105
6.2 *Multi-DLMPVis*: An interactive dashboard for multiple deep learning models performance visualizations (Figure courtesy [8]). 106
6.3 An interactive deep visual analytics system which is consisted of four major parts. A) distribution view, B) demographic chart, C) patient history, and D) knowledge graph view (Figure courtesy [166]). 110
6.4 The architecture of *UCBVis* system. 117
6.5 System architecture of *Multi-DLMPVis* System. 119
6.6 The input selection and performance ranking view of the *Multi-DLMPVis* System. 122
6.7 3D projection, confusion matrix, and misclassified instances view of the *Multi-DLMPVis* System. 123

LIST OF TABLES

TABLE NAME **Page**

2.1 List of the key journals and conferences. 13

3.1 Key studies: different methods for adverse-selection detection in the insurance market. 27
3.2 Characteristics of demographic dataset. 34
3.3 Different interestingness measure of the different patterns. 35
3.4 The result of the experiment. 38
3.5 Advantages and limitations ML/DL techniques. 42

4.1 Characteristics of respondents 57
4.2 Conditional probability table (CPT) for bayesian network (BN). 58
4.3 Dataset description. 62
4.4 Diabetes dataset description. 63
4.5 Characteristics of respondents. 65
4.6 Feature comparison. 68
4.7 User study results. 75

5.1 Key questions identified in collaboration with domain experts. 89
5.2 Types of queries and visualization observed in this study. 94
5.3 User study response for output visualization to post-study questionnaires. . . 97
5.4 Key observations identified in collaboration with domain experts where expert responses 1, 2, 3, 4, and 5 indicate strongly disagree, disagree, neutral, agree, and strongly agree, respectively. 98

6.1 Overview of key representative works in visual and deep visual analytics. . . 111
6.2 Overview of various evaluation techniques and their applications. 115
6.3 Demographic characteristics of claim data set. 116
6.4 Domain specific requirements. 117

6.5 User study results. 124
6.6 Feature comparison analysis. 125

Abstract

Customer behavior refers to the study of customers and the procedure they use to pick, use, and dispose of products or services. The understanding of customer behavior analysis (CBA) is essential for improving business strategies. The existing studies have explored useful information to analyze customers' behaviors. However, they often fail to allow the analysts, including business management, development, decision-making, etc. Notably, the existing research on CBA is limited with four main challenges. First, the analysis of the absence of useful private information and the presence of asymmetric information of customers, e.g., discover adverse information in each cell rather than for each data instance. Second, exploring customer behavior with multi-dimensional and temporal data is necessary for any competitive and global business to improve its strategies. Third, the estimation of the correlation between claim analysis and risk management is key to avoiding fraud; Fourth, the lack of quantitative research necessitates performance analysis at the class, instance levels, and model visualization. Several approaches to addressing these issues were introduced that are inconsistent with models of rational choice. Due to the excellent ability to collect and classify valuable knowledge, data mining has become a standard support method for gaining interesting insight into customer behavior. Even though rapid and accurate identification of customer demands is critical to business management, it is not feasible to design all approaches to meet all criteria to be developed. Therefore, this thesis aims to exploit novel data mining techniques blending with visual analytics (VA) to explore customer behavior and provide valuable insight for decision-making support. Insurance data such as questionnaires, demographic, and claim data are used as a testbed to demonstrate our techniques. This thesis is categorized into four main themes: (1) pattern mining (PM) for discovering adverse behavior (AB); (2) visual analytics (VA) for exploring customer behavior; (3) natural language interaction driven data visualization (NLI-driven-DV) to analyze customer claim behavior and manage risk; (4) deep visual analytics (DVA) to provide a wide range of performance evaluations of different methods for understanding customer behavior (UCB). This is one of the first studies to utilize data mining techniques blending with visual analytics (VA) for exploring customer behavior from the insurance business aspect. The empirical results of this thesis show the advantages and effectiveness of the developed methods valuable for researchers and insurance managers (IMs). Moreover, various aspects of insurance data have been researched and integrated into sophisticated visual interactive systems (VIS) to gain a deeper understanding of customer behavior and to better business plans and make decisions.

Keywords: Behavior analysis, data analytics, data visualization, visual analytics, pattern mining, machine learning, risk management, business strategies.

List of Abbreviations

CBA	Customer Behaviour Analysis
AB	Adverse Behaviour
AS	Adverse Selection
IMs	Insurance Managers
PM	Pattern Mining
ASF	Adverse Selected Factors
BFP	Breaking Frequent Patterns
SEIFA	Socio Economic Indexes for Areas
RFA	Randomly Flipping Attribute
MHI	Mental Health Illness
MHIVis	Visual Analytics for Exploring Mental Health Illness
UCB	Understanding Customer Behaviour
UCBVis	Visual Interactive System for Understanding Customer Behavior
VA	Visual Analytics
VAS	Visual Analytics System
VIS/Vis	Visual Interactive System
D3	Data-Driven Documents
ExVis	Explainable Visual Interactive System
DAG	Directed Acyclic Graph
BN	Bayesian Networks
CPT	Conditional Probability Table
KLD	Kullback-Leibler Divergence
JSD	Jensen-Shannon Divergence
AI	Artificial Intelligence

ML	Machine Learning
DL	Deep Learning
ExTraVis	Exploration of Traffic Incidents Using a Visual Interactive System
ITD	Incident Trend Dashboard
FARS	Fatality Analysis Reporting System
FTS	Free Text Search
TIC	Traffic Incident Controllers
NLP	Natural Language Processing
NLQ	Natural Language Query
ExNLQVis	Explainable NLQ based Visual Interactive System
DVA	Deep Visual Analytics
DS	Data Science
IV	Information Visualization
LIWC	Linguistic Inquiry and Word Count
LR	Logistic Regression
LSTM	Long Short Term Memory
UI	User Interface
HCI	Human Computer Interaction
RM	Risk Management
DM	Diabetes Mellitus
DiaVis	Visual Interactive System for Exploring Diabetes Disease
ARM	Association Rule Mining
CRM	Customer Risk Management

Chapter 1

Introduction

This chapter presents the background and motivation of this research in Section 1.1. Section 1.2 discusses the thesis objectives followed by the thesis limitations and contributions in Section 1.3 and 1.4. The organization of this thesis is outlined in Section 1.5.

1.1 Background and Motivation

Customer behavior analysis (CBA) means the study of individuals, groups, or organizations about their process of securing, selecting, using, and disposing of the services, products, experiences, or ideas to satisfy needs and the impact of these processes to the society [40, 124]. It has been increasingly highlighted in many fields, such as web search and usage [5], customer relationship management, the insurance industry, government, the financial services industry, and so on [120]. The current trend on CBA has been recognized on the business problem rather than on the information technology (IT). In business, the study of CBA is a significant factor determining customers' intent to purchase or not to purchase products or services. Therefore, the motivation of this thesis is presented using the following scenarios in analyzing customer behavior.

1.1.1 Analytics for Decision Making

Understanding customer behavior (UCB) is the key to effective business planning, which is crucial to the success of the business industry. For instance, customers can intentionally provide misleading information to the insurer to avoid paying higher premiums in the insurance context. Also, they may give negative feedback for the service criteria. An insight into how standards interact in guiding customers' intention can provide business managers with insight into the preferences. However, understanding CBA is a complex process, usually involving several factors depending on their preferences and the nature of data. For instance, the missing useful private information, asymmetric information, anomalous information at the cell level rather than the data instance level, and a lack of quantitative research [120]. Let's consider an example to illustrate this issue. An adverse selection (AS) occurs when a policyholder obtains a policy at a much lower premium than the insurance company would charge if they were aware of the actual risk regarding the applicant, usually because the applicant withholds relevant information or provides false information that thwarts the effectiveness of the insurance company's risk evaluation system. Over the past few decades, researchers have tried to help the business strategy by clearly understanding customer behavior (UCB). It is evident from the existing research that few studies related to CBA have been carried out. Therefore, if a novel and complex decision-making process can be modeled, the customer's preference can be revealed and analyzed effectively.

1.1.2 Customer Behavior Discovery

Understanding customer behavior (UCB) is essential in Australia's life insurance industry, aiming to provide more insightful information that improves business operations and decision-making. This means insurance authorities such as insurance managers (IMs) can gain insights into customers, such as how customers claimed at their original state or suburb in Australia, how frequently they claimed, etc. However, IMs often cannot collect the customer's exact information because of the lack of proper understanding and computational expertise. Thus, it is becoming a severe economic problem. For example, away from less interest for authority budgets, claim-cause evasion has contributed to a bias business race environment for consistent policyholders whose working expenses are higher than rebellious contenders. To make solid business strategic, IMs seek to understand customer behaviors. Therefore, this study attempts to (i) enquire the domain requirements for understanding customer claim behavior sequences, (ii) understand whether customer claim behavior influence in the life insurance, and (iii) investigate the importance of the precise formal and informal information causes from which individuals get financial knowledge.

1.1.3 Customer Behavior Visualization

Visualizing customer behavior is an essential part of defining the most successful business plan and, as a result, a crucial part of developing a business strategy [124]. It involves an individual or an organization and the procedures for determining the necessary products, services, or ideas to meet customer demands and their implications for purchasing habits. Organizations combine psychological, social, and anthropological components of customer behaviors to determine the most effective approach to customer strategic positioning to develop a proactive determination of customer behaviors. For example, Wassouf et al. [287] and Kalaivani and Sumathi [132] introduced a new wave of customer relationship management strategies using extensive data analysis, which described customer behavior and understanding their needs. They proved that their proposed FBPCA model provides the best accuracy to the existing models. Additionally, emotions and personality are essential factors in decision-making because they influence how we communicate with others [66]. As a result, a better understanding of emotional capacity can help us better understand customer behavior visualization. For example, Halkiopoulos et al. [98] proposed how emotional processing influences purchasing decisions, which choices consumers with a high emotion intelligence (EI) make more readily than those with a low EI, and how EI might affect relationships between key customer variables such as impulsivity and purchase intention. Although current studies focused on traditional visualization techniques, it has been observed that visual analytics (VA) for customer relationship management turned into a significant sector where various models applied to analyze customer's social data and solved diverse challenges such as (i) vast amounts of dataset, (ii) factor-based prediction technique, (iii) progressive data-driven approach, and (iv) data mining technique to discover patterns. To address these issues, Khade [138] applied visualization techniques for performing CBA to explore the biggest challenge of identifying confidential information through the vast amount of data. Islam et al. [124] presented a pattern mining-based system to analyze customer behavior and described significant challenges by applying visualization technique. They exemplified the insights of customer behaviors and explained how to gain behavioral understanding from insurance data.

1.1.4 Summary

The above discussion has illustrated several essential issues in CBA in an insurance context. IMs and researchers desire insightful knowledge of decision-making, customer adverse behavior (CAB) identification, and claim behavior preference for effective business management. However, very few studies have been fully satisfied because existing systems and techniques cannot perform the analysis tasks effectively. There are few approaches for capturing the complex decision-making process of the customer to the actual scenario in decision-making process analysis. Additionally, no method has been reported to capture and identify the CAB successfully. No interactive visualization system (VIS) was written for claim behavior analysis and risk management to effectively visualize the customer claim records from the abundant claim data resources. If these research gaps are overcome, it will be easier to support CBA duties and better understand client preferences. As a result, effective strategic planning and decision-making can help businesses enhance their performance.

1.2 Research Objectives

This thesis intends to improve the efficiency of CBA tasks by establishing new methods and methodologies for better modeling customers' decision-making processes, evaluating their behavior more efficiently, and identifying their emerging preferences more effectively. Furthermore, the issues of evaluating and interpreting customer behavior for decision-making, business planning, and customer activities are also addressed. Thus, the following research goals are the main focus of this thesis:

- To analyze the life insurance policyholder's behavior to identify adverse behavior.
- To design an interactive visualization system to provide deeper insight into policyholders' mental health states to help business management controllers make decisions.
- Designing and implementing a novel VIS with essential data to provide insightful evidence and explain system results.
- To investigate the impact of demographic variables and explore customer health conditions and their associated factors.
- To design visual analytics system enhanced by NLIs for risk management and claim analysis.
- To reflect on assessment in the visualization system through a systematic understanding of numerous methods.

1.3 Research Problems

Insurance industries are often uncertain because of the actions imposed by the unregulated movement of high-risk policyholders [216]. The authorities need to remain alert for changes, demands, and necessary steps to manage and work with public users [33, 125, 136]. As a result, the authorities are keen to understand customer behavior to develop appropriate business strategies. However, it is very challenging to identify them with some hidden characters related to different data of the insurance policy. Therefore, IMs need to have access to all the critical aspects of the related information, detailing which packages are the best to promote a product as follows:

- What are the common factors to continue towards adverse behavior in insurance domain?
- How to build an unsupervised machine learning model to detect potential adverse behavior of customers?

The exploration of customer behavior is essential in the life insurance industry in Australia, aiming to provide more insightful information's improving business operations and decision-makings [144, 260]. However, because of the lack of proper understanding and computational expertise, IMs often cannot collect the customer's exact information. Thus, it is becoming a severe economic problem [167]. To make solid business strategic, IMs seek to understand customer behaviors [77]. Therefore, this study attempts to enquire the domain requirements for exploring and understanding customer behavior sequences as follows:

- Whether customer behavior influence in life insurance?
- How to investigate the importance of the precise formal and informal information causes from which individuals get financial knowledge?
- How to visualize the policyholder's mental health behaviors and why certain recommended information remains open problems?
- How explainable visual decision support system could help financial services thrive?

Customer claims behavior analysis for risk management is crucial to avoiding fraud and managing risk in the life insurance industry. Though analyzing claims behavior plays a fundamental role in supporting analysis tasks in the business domain, interactively visualizing user behavior remains a challenging task. Therefore, this study attempts to visualize and monitor the policyholders' claim risk based on the questions as follows:

- Why is an interactive visual analytic tool necessary for the insurance domain?
- What the key factors/content should be depicted when exploring insurance claims and visualizing risks/risk-related information?
- Who will our visualization system monitor to control risk?
- When stakeholders should consider an interactive risk visualization system a useful tool in light of the benefits it provides?
- How can natural language interfaces (NLIs) be supported through an interactive visualization system to investigate insurance claims and manage risk?

Visual interactive system (VIS) has been received significant attention for solving various complex problems. However, designing and implementing a novel VIS with a large scale of data is a challenging task. While existing studies have applied various visual analytics (VA) to analyze and visualize insightful information, there are still enough spaces to explore CBA [117]. Therefore, this thesis aims to address the following questions:

- How to design and implement a novel VIS using deep learning (DL) techniques to analyze and visualize insightful information for CBA?
- Why should various aspects of a DL model be visualized?
- What and how to visualize in DL?
- When will the visualization phase take place during the process of developing and training a network?

While most of these tasks seem simple to humans, they are extremely difficult for computer algorithms to solve because there is no systematic explanation of how to solve them.

1.4 Research Contributions

The key contributions of the thesis are:

The first of the research questions was to analyze how to identify the adverse behavior (AB) of the policyholders in the life insurance industry? An extensive questionnaire-based behavioral dataset from one of the most popular insurance companies in Australia covering about $31,870$ data records of the different policyholders and includes 834 columns, each about a yes or no question. Additionally, the demographic dataset contains information on the policyholders' ID, gender, postcode, age, and occupation. In addition to the original data, I created a synthetic AS dataset by randomly flipping the attribute values of 10% of the records in the test set. A novel association rule learning-based approach 'ARLAS' is proposed to detect the AS behavior of policyholders. The experiment results on $31,800$ policyholders show that the proposed approach achieves significant gains in performance comparatively [120].

The second theme dealt with visualizing consumer behavior using multi-dimensional and temporal data to deliver new insights and better business plans for any competitive and global business. To allow the analysts, including business management, development, and decision-making, a data-driven visual decision support system such as (i) *ExVis* - explainable visual interactive system for risk management, (ii) *MHIVis* - exploring mental illness of customer with visual interactive system, and (iii) *DiaVis* - an interactive visual system (Vis) to explore diabetes mellitus (DM) insights and its associated factors are design. A large number of customer behavioral records are used to facilitate the exploration of CBA. The robustness of these systems through a user study with five participants shows that *ExVis*, *MHIVis* and *DiaVis* are perceived to be more practical and provide actionable insights [121, 126, 213].

The third theme was related to exploring customer claim behavior for risk management in the life insurance industry. A visual analytics tool enhanced by NLIs for risk management and claim analysis is designed to support business analysts. I performed a user study of 10 experts to evaluate its performance, which suggests that our visualization system can provide better insights and assistance to insurance managers (IMs) in reducing loss and guiding changes to insurance premiums policies. Furthermore, I provide a concise set of guidelines to visualize risk to avoid dangers in the insurance domain. I also discuss the challenges associated with using a visualization system in the insurance industry, focusing on aspects related to visualization research [116].

The final theme was designing and implementing a novel VIS using deep learning techniques with a large scale of data to analyze and visualize insightful information. Thus, I consider two parts of design methodology as a dashboard to bridge the information gap through visual representation and interaction techniques. First, I present the customer behavior pattern of multi-dimensional relationship through the visualization system named *UCBVis* based on interweaving the pattern mining and querying with a designed encoding scheme [124]. After that, I demonstrate a visual application named *Multi-DLMPVis* for multiple deep learning models performance visualizations, where I apply five DL models such as CNN, VGG16, AlexNet, DenseNet, and ResNet50. To establish our *UCBVis* visualization system, I use a large number of customer claim records and present visual outcomes to facilitate the exploration of customer behavior. However, to establish our *Multi-DLMPVis* system, I consider an image dataset, which is publicly available where I will fit our customer records in future work [146]. Thus, my target is to reflect on assessment in the visualization system through a systematic understanding of numerous evaluations.

1.5 Thesis Organization

The rest of the thesis is organized as follows:

Chapter 2 presents a detailed description of the understanding of CBA, visual analytics in CBA, and data mining techniques. In the subsequent chapters of this thesis, I emphasize the current and emerging issues in the insurance industry, which defines the theme of technological development.

Chapter 3 presents a novel association rule learning-based approach 'ARLAS' for discovering the adverse behavior of the customer. I briefly describe this approach to locate repeating relationships between unique items in a data set and represent them in the form of association rules.

Chapter 4 introduces visual analytics for analyzing CBA for supporting insurance management tasks. Our design study addresses a known problem with a novel solution and provides data-driven visual decision support in collective policy data.

Chapter 5 proposes natural language interactions-driven data visualization (NLI-driven-DV) methods for exploring customer claim behavior and managing risk. Two integrated techniques are briefly described, including the natural language interaction for query processing and the visual analytics for customer behavior representation.

Chapter 6 proposes pattern mining (PM) and deep learning (DL) with an interactive visualization system for understanding CBA. These methods are frequent pattern mining, convolution neural network (CNN), VGG-16, AlexNet, ResNet-50, and DenseNet approaches. All these methods are integrated with VIS to understand more about CBA.

Chapter 7 presents the conclusion by summarizing both theoretical and practical contributions and suggesting some possible directions for future research.

CHAPTER 2

Related Work

This chapter provides a deeper look at customer behavior analysis (CBA) and quantifies these behaviors within the insurance industry using advanced data mining techniques. To determine the novel scopes of applying data mining techniques, it is essential first to identify what has been done so far, what could have been done, and the limitations of the existing studies. Therefore, this chapter has mainly been carried out from four perspectives: 2.1. literature search methodology; 2.2. understanding customer behavior (UCB); 2.3. visualizing customer behavior (VCB); 2.4. data mining techniques for customer behavior analysis (DM for CBA); and 2.5. summary.

2.1 Literature Search Methodology

This section provides 'literature search methodology' research efforts where I briefly discuss how I drew in our work. As shown in Figure 2.1, I provide the keyword searching was conducted across four electronic databases to locate relevant papers from diverse publishers. For example, the terms 'customer behavior analysis', 'visual analytics', 'visualizing customer behavior', 'data mining for customer behavior analysis' and 'evaluating visual analytics' were searched in the ieee explore, acm, science direct, arXiv, wiley, and google scholar database to satisfy PRISMA criteria [201]. The terms were defined subjectively and applied to each database to obtain the most significant volume of relevant articles. However, articles that were not published in english and book chapters, newspaper articles, unpublished articles, and non-scientific articles were not excluded.

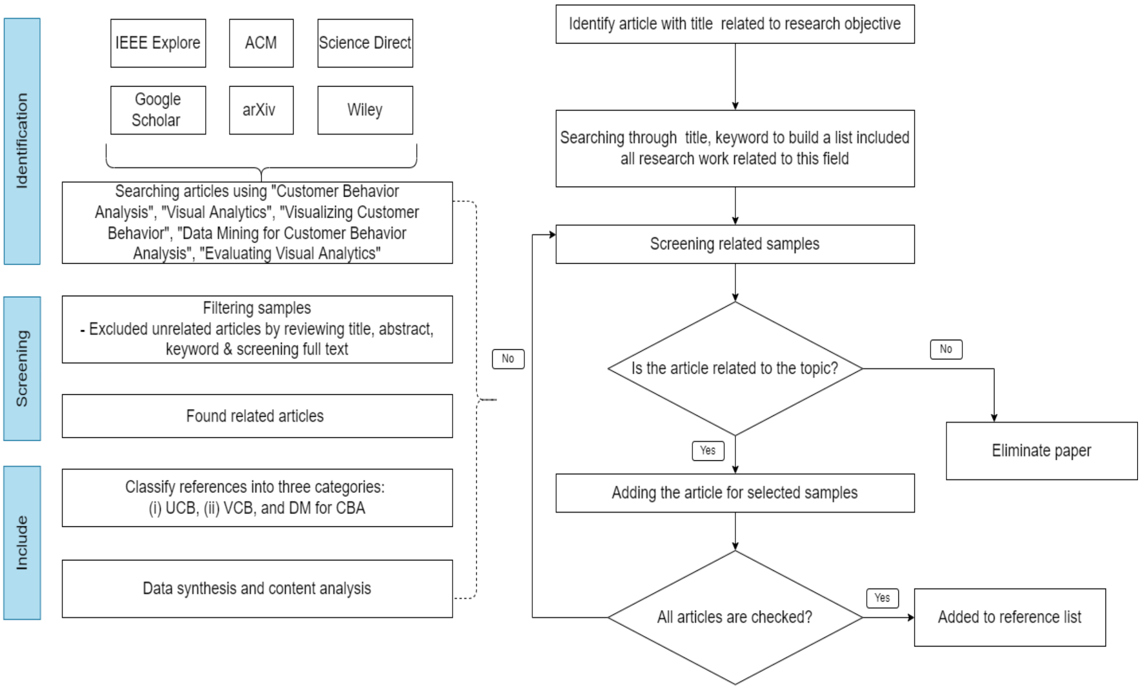

Identification
IEEE Explore
ACM
Science Direct
Google Scholar
arXiv
Wiley
Searching articles using "Customer Behavior Analysis", "Visual Analytics", "Visualizing Customer Behavior", "Data Mining for Customer Behavior Analysis", "Evaluating Visual Analytics"
Screening
Filtering samples
- Excluded unrelated articles by reviewing title, abstract, keyword & screening full text
Found related articles
Include
Classify references into three categories: (i) UCB, (ii) VCB, and DM for CBA
Data synthesis and content analysis
Identify article with title related to research objective
Searching through title, keyword to build a list included all research work related to this field
Screening related samples
Is the article related to the topic?
No
Eliminate paper
Yes
Adding the article for selected samples
All articles are checked?
No
Yes
Added to reference list

Figure 2.1: Proposed review methodology for sample collection and analysis.

This literature review mostly examines articles published in top-tier journals and conferences from 2014 to 2021. According to Australian Business Deans Councils (ABDC) and The Computing Research and Education Association of Australasia (CORE), the study focuses on top-tier journals and conferences, respectively, because there is no standard list of research journals and conferences in this discipline. The key journal and conference names covered in this review are included in Table 2.1. Most of these journals and conferences are ranked A/A*, which the Australian Business Deans Councils (ABDC) and CORE proposes. In addition, some tier B journals and conferences are also included, as they are highly cited by data science researchers [124, 138]. These journals and conferences are “Expert System with Applications," and “International Conference on Information Visualization". Thus, the following section examines the content of these research articles to provide an overview of CBA research themes and related data mining approaches.

2.2 Understanding Customer Behavior

This section focuses on understanding customer behavior (UCB) and how stakeholders solve problems using data mining and visual analytics. I discuss two main parts of the

Table 2.1: List of the key journals and conferences.

Journal Name	Ranking (ABDC 2020)
The Quarterly Journal of Economics	A*
Technological Forecasting & Social Change	A
Decision Support System	A*
Annals of Operation Research	A
Journal of Risk and Insurance	A
Expert System with Applications	B
Information Management	A*
IEEE Transactions on Visualization and Computer Graphics	A*
Australasian Journal of Information Systems	A
IEEE Transactions on Intelligent Transportation System	A
MIS Quaterly	A*
Knowledge-Based Systems	A*
Conference Name	Ranking (CORE 2020)
ACM International Conference on Research and Development in Information Retrieval	A*
Visual Analytics Sciences and Technology	A
AAAI Conference on Artificial Intelligence	A*
CHI Conference on Human Factors in Computing Systems	A*
Information Visualization	B

work that relate to this problem. First, I set the categorization to articulate the CBA. Second, I identify the current research issues by outlining existing research efforts to communicate the CBA in the life insurance industry.

2.2.1 Categorization of Customer Behavior

This section has mainly been carried out from three perspectives: 1) understanding customer adverse behavior (UCAB); 2) understanding customer claim behavior (UCCB); and 3) understanding customer mental health behavior (UCMHB).

2.2.1.1 Customer Adverse Behavior

First, adverse behavior, also called adverse-selection (AS) behavior from customers is typical and presents a risk to the integrity of the insurance market [69]. This is where high-risk customers can intentionally provide misleading information to the insurer to avoid paying higher premiums or to avoid being excluded for eligibility [96]. Private asymmetric insurance market information has been claimed for AS in various studies in which the policyholders are better informed about the distribution, likelihood, and use this information to select their insurance plans and size of risk or losses. Moreover, several insurance market studies have extensively highlighted the potential importance of asymmetric information, customer engagement, and financial knowledge and documented various responsible factors for customers' AS behavior for insurance industry development and sustainability [61]. For instance, Puelz and Snow [209] provide a piece of evidence for AS in US automobile insurance markets, but only for experienced drivers. Cohen and Siegelman [61] reviewed a large number of the empirical studies and insisted that evidence of AS on health insurance markets was found in a considerable collection of empirical studies, people in poorer health prefer policies that provide more generous coverage, or policyholders who buy more insurance coverage appear to be riskier. Although engagement classification is mainly linked to AS in life insurance markets, there is no engagement classification for observing or measuring what individual factors are more likely to AS, which individual customers are engaged, disengaged, and those in between. As a result, it is impossible to identify accurate AS users, and many honest policyholders may suffer. Existing research use customer behavior attributes data-however, demographic and socio-economic information, which has not received much research attention for AS purposes. Moreover, the AS hypothesis test is usually performed using bivariate or multivariate models, generating an endogeneity bias in the estimation results. That implies that by considering a new modeling framework, an insurer can prevent AS and generate income.

2.2.1.2 Customer Claim Behavior

Second, customer claim behavior analysis has been highlighted in several areas, including web search and usage, customer relationship management, government, the financial services industry, and so on. Many studies have been undertaken to utilize the data generated from insurance customer claim behavior. For example, an earlier study by Kim et al. [140] detected the various changes in customer behavior and performed poorly by exhibiting a lack of understanding of fundamental ideas associated with customer claim behavior issues. Although customer claim behavior has been the topic of several investigations and this concept is considered well documents, very few published research has been utilized to deeply understand customer claim behavior [49, 67, 295]. Most of the existing literature has been focusing more on anomalous information using customer demographic data. However, very few customer analytics research has been conducted using customer claim data within the life insurance industry. Additionally, existing researchers have also been attempted to incorporate customer claim behavior using some traditional models. However, existing methods can only solve formal client behavior analysis based on static customer attributes but not sequential behavior of customer claims. Only a few research has attempted to analyze and identify practical solutions for the insurance company regarding the customer claim behavior analysis. To measure customer claim behavior, I mainly look at the customer sequential claims behavior variables to an extra deeper insight into knowledge in the insurance industry. Although there are few kinds of literature from the business side tried to capture these problems using survey and other qualitative methods [67]. From a data science perspective, these can not visualized fully as decision-making problems. Therefore, visual analytics could help to understand the current state of the art, give a deeper look at consumer behavior, and quantify these behaviors within the Australian life insurance industry.

2.2.1.3 Customer Mental Health Behavior

Third, there has been a dramatic increase in research on the relationship between mental health disorders and claim management in recent years. Several studies have been shown that mental illness leads to an increase in the amount of sickness absence (SA) from work [72, 101]. For example, employed individuals with mental illness have three times higher SA rates from work than individuals who do have a mental illness [219]. The Norwegian Disability Pension Registry (NDPR) analysis showed that mental disorders were responsible for the most working time lost over two years [142]. Estimates from the UK indicate that 40% of all active time can be attributed to mental health disorders [288]. The OECD reported that a third of all disability pensions awarded in European countries were mental health-related [219]. In 2001-2014, the number of disability pensions awarded for mental health conditions rose 50% in Australia. Therefore, I spend attention on the customer's mental health condition for predicting claim. It is noted that insurance companies provide safety to society by offering financial risk insurance. Transferring the risk to the insurer in exchange for a fixed premium allows individuals to trade uncertainty to certainty. An insurer sets the price for insurance before the actual cost is reported. Because of this phenomenon, known as the insurance business reverse production cycle, it is important for an insurer to correctly assess the risks in their portfolio. To this end, predictive modeling tools come in handy. Many studies have pointed out the risk classification method and predicted future claims' frequency and loss severity. For instance, Wuthrich et al. [293] show how tree-based machine learning techniques can be adapted to model claim frequencies. Pesantez-Narvaez et al. [205]

employ XGBoost to predict the occurrence of claims using telematics data. Ferrario et al. [87] and Schelldorfer and Schelldorfer et al. [236] propose neural networks to model claim frequency directly or via a nested GLM. On the other hand, although previous studies have shown some traditional machine learning techniques in the insurance field, very few studies have considered advanced machine learning models, making them hard to meet the practical requirements. Very few customer analytics research has been conducted using customer mental health condition data within the life insurance industry. Moreover, the predictive performance of the existing techniques tends to be relatively low. Therefore, I conclude that there is a novel scope of applying advanced machine learning and data analytics methods for analyzing customer mental health data analysis.

2.2.2 Current Customer Behavior Issues

The above section provides an overview of CBA categorization in the context of the insurance sector. This section looks into several CBA issues to identify the obstacles IMs face and current approaches' limitations. Customer adverse behavior, customer claim behavior, and customer mental health analysis are highlighted as essential aspects of insurance management. As our research is focused on a local insurance company in Australia, some information on the present state of the Australian insurance sector is provided accordingly.

2.2.2.1 Issues in Customer Adverse Behavior

Adverse selection (AS), as it is also known, refers to a situation where sellers have information that buyers do not have, or vice versa [61]. In the Australian life insurance business, AS behavior from customers is typical and presents a risk to the integrity of the insurance market [196]. This is where high-risk customers can intentionally obscure or provide misleading information to the insurer to avoid paying higher premiums or to avoid being excluded for eligibility. For the Australian life insurance market, AS is one of the significant sources of market failure [120]. For example, a race car driver may obtain a life insurance policy without informing that he has a dangerous occupation. In another example, a user may get insurance coverage providing a residence address in an area with a very low crime rate when the applicant lives in an area with a very high crime rate. As a result, insurance companies often have to get losers. Therefore, the ability to detect AS in the insurance market is critical to reducing company losses, increasing service quality, improving risk adjustments when assessing AB, allowing insurers to focus on complications they have the most concern for, and allowing for the improvement of insurance premium policy.

2.2.2.2 Issues in Customer Claim Behavior

The behavior of customer claim plays a vital role as an internal driving force for many business issues [260]. Hence, a deeper understanding of customer behavior is crucial to provide customized services to specific groups of customers [56]. Roughly speaking, the good customers are those paying expensive premiums but having few claims over time. On the other hand, fraud customers are those paying fewer premiums but having more claims [104, 223]. Life insurance, in particular, is the most critical potential risk transfer tool. Incorporating claim behavior analysis into an assessment of the economic value of a customer is an important conceptual development that has both theoretical and practical implications [10, 124]. The claim of an existing customer can originate from various factors that affect the volatility of future revenue. Traditional behavioral analysis of customer claim is usually focused on static customer attributes, i.e., demographic data. However, sequential behavior, which has not received much research attention to date, is much more valuable. Without past customer relationships, the risk of undesirable behaviors such as personal characteristics, past claims history, and related customer demographic information is more prominent. Meanwhile, it is challenging for the insurance industry to forecast such risk because of the limited information. So, analyzing the scope of the customer claim behavior and its detection is essential for making business fair.

2.2.2.3 Issues in Customer Mental Health Behavior

Mental health conditions (MHC) are a growing reason for claiming injury compensation in Australia; however, very few is known about how various factors handle these claims to injury entitlements [101]. Additionally, MHC claims are the most expensive and challenging to manage of all the Australian claims categories [236]. Sickness absence, depression, anxiety, mental stress, etc., can harm the physical and psychological health and productivity of an individual's work. The literature shows a clear correlation between MHC claims, long-term work absence from work, and high disability pension rates [142]. It is necessary to first consider the existing obstacles and facilitators to return to work (RTW) for individuals with either MHC claims and/or mental disorder to improve mental health outcomes, minimize costs associated with MHC claims and promote the timely return to work (RTW) [288]. However, gold-standard labels in MHC contain noisy information on the psychiatric mental health status of patients. While a growing research body combines insurance data with machine learning techniques to predict individual mental health conditions, it is challenging to deal with noisy labels from sensitive patient populations.

2.2.3 Summary

In summary, the above studies focusing on CBAs can be broadly divided into three categories: (i) adverse-selection behavior discovery, (ii) claim behavior analysis (CB), and (iii) customer mental health condition identification. Existing literature mainly addresses the issues in analyzing customer outlier behavior and predicting mental health claims management. Concerning the current literature within the insurance CBAs field, multiple notable research initiatives have been working on this topic within Australia. These studies have shown some challenges in the current CBAs system, which makes the outcomes insufficient. Recent research suggests the critical challenges for the insurance company are:

- Absence of useful private information and presence of asymmetric information of customers.

- Detect anomalous information in each cell rather than for each data instance.

- Estimation of the correlation between risk and coverage.

- Noisy information in labels e.g. many of the claims are just by chance.

- Lack of quantitative research in this field.

Additionally, the authors have incorporated a good number of machine learning and data analytics techniques in these studies. However, I argue that there still exists a new window of opportunities to contribute in this field by incorporating advanced machine learning and data analytics techniques as existing studies do not promote highly individualized customer behaviors and demand management in the insurance industry. These research gaps are the primary motivation for this thesis and related customer behavior analysis research.

2.3 Visualizing Customer Behavior

To analyze data and identify its very subtle trends, patterns or context in graphical format, data visualization (DV) is a handy tool in this regard. Replacing boring tables, charts, or graphs, DV enables users to represent data in a very compact and interactive way through maps, bars, pie, etc. Nowadays, it is used to display more complex data, identify the risk, and make a decision exploiting the very cognitive and communication power of human beings in security, economy, healthcare, etc. For example, complex clinical data set are observed and analyzed by psychologists with a DV approach for child developmental health [120].

DV is now extensively used in financial and business institutions as well. For instance, Eppler and Aeschimann [85] proposed some solid guidelines and frameworks for careful operation. Huang et al. [112] proposed a visual analytical framework to prevent fraud attacks and to assist the human cognitive process. Similarly, Leite et al. [160] investigated to avoid any harmful transactions through a visual analytic approach to integrate human analysis into this process. Their approach focuses on uncovering fraud operations identifying unusual financial events, and fine-tuning existing automatic fraud alert systems. Moreover, networked guarantee loan risk management examines groups of enterprises who support each other for financial security [197]. Therefore, to build healthy communication between risk experts and decision-makers, an interactive visualization could be an efficient solution to identify, understand and analyze the most relevant risks [85, 112, 160, 197].

2.4 Customer Behavior Analysis using Data Mining Techniques

The analysis of consumer behavior described in Section 2.2 and 2.3 is conducted using data mining techniques. Each activity can be accomplished using strategies for discovering hidden customer behavior in the data. Data mining techniques have begun as a nontrivial process of detecting legitimate, innovative, potentially helpful, and eventually intelligible patterns in structured databases to address the requirement to discover hidden knowledge from vast amounts of data [37, 192]. For example, Hohman et al. [107] defined data mining as the process of extracting and identifying relevant information from big databases using statistical, mathematics, artificial intelligence, and machine-learning approaches. Islam et al. [120] provides a similar definition of data mining as the practice of extracting or detecting 147 hidden patterns or information from massive databases. Thus, data mining technologies are the most popular supporting tool for business decision-making since they excel at extracting and recognizing important information and knowledge from extensive databases. The following section describes the categorization of DM techniques for conducting the CBA research in the insurance context.

Association Analysis: Association analysis is mainly used for market basket analysis to identify the frequent presence of the combination of items purchased together [6]. The most frequently used algorithm is Apriori for association rule mining in an insurance context [120, 129, 192].

Clustering: The task of segmenting a diverse population into multiple homogeneous clusters is known as clustering [31, 169]. Clustering is an unsupervised process in which the clusters are unknown when the algorithms are first to run. K-means clustering and hierarchical cluster analysis are two of the most well-known cluster analysis methods [31].

Predictive Modeling: There are two main types of predictive modeling techniques: classification and regression. Classifying records based on specified features aims to build a model for predicting future customer behavior [194]. For example, the support vector machine is the most common technique for classification [307]. On the other hand, regression is a statistical estimation technique that maps each data record to an actual value. The most often utilized techniques, for example, are linear regression and logistic regression [262]

Forecasting: Forecasting is a method of predicting the future based on patterns of records that deal with constantly valued outcomes, as described in [111]. It has to do with modeling and its logical relationship in the future. A forecasting application like the demand forecast is a good example [82].

Visualization: Visualization refers to the presentation of data so that users can view complex patterns [117, 280]. It is often used in conjunction with other data mining models to provide a deeper understanding of the relationships [78]. Moreover, some of the tools for visualization are various depending on the analysis used in financial issues [147, 160]. The two most common types of data visualization are static and interactive. Static visualizations are nothing more than a single but showing an informative view of a particular data story. On the other hand, interactive visualization uses graphic representations of data to define or explain how we engage with data.

Statistical Analysis: Statistical analysis refers to various statistical analysis techniques used to create a variable analysis and hypothesis testing. For example, analysis of variance (ANOVA), multivariate analysis of variance (MANOVA), factor analysis, structural equation modeling (SEM), correlation analysis, and some form of regression analysis are all popular techniques in the customer behavior analysis (CBA) literature [242].

Furthermore, several data mining approaches are frequently used to support, forecast, or validate the impact of a corporate plan. SPSS, for example, can be employed as a statistical learning tool or Tableau as a visualization tool. Logistic regression can be used as a classifier to categorize customer profiles or as a statistical research technique to investigate the relationships between variables. It can also be used as a predictor in multiple linear regression and as a statistical learning method.

2.5 Summary

This chapter has reviewed the literature search methodology about data mining and visual analytics techniques for business risk management and identified several potential research issues in the insurance domain. It provides various applications of DM techniques in CBA, which has received increasing attention from researchers. For example, DM techniques such as pattern mining, clustering, logistic regression, neural networks, linear regression, and Statistical methods such as ANOVA, MA, etc., have been used widely by data science and business researchers. In addition, association rule minings are frequently used in analyzing customer behavior preferences. Besides, the CBA is critical to decision-makers in business development, planning, and risk management. Claim benefit information has appeared as a new dimension to capturing CBA comprehensively and efficiently. However, the claim data is noisy and misleading, which requires novel techniques to reveal confidential customer behaviors. There is a demand for a better method to solve the actual situation. Furthermore, IMs need to visualize the emerging claim preference of policyholders. Traditional methodologies are unable to capture the growing pattern in client data fully. Therefore, new methods are required to meet evolving preference analysis issues better. Furthermore, this thesis aims to develop unique data mining techniques for addressing CBA difficulties in the insurance business. In Chapter 3, the difficulty in customer AB is addressed by proposing a new approach based on pattern mining. It beats conventional AB approaches because it can assess several criteria simultaneously and account for all interactions between them. In Chapter 4, a new visual analytics framework is used to represent the problem of detecting customer behaviors. This system can process mental health data effectively to provide preferred IMs attractions. The IMs employing customer claim data in Chapter 5 will benefit from the natural language supporting VIS. In Chapter 6, the development of a new VIS based on the emerging pattern mining (PM) and deep learning (DL) approach is presented as a solution to the CBA dilemma.

CHAPTER 3

Pattern Mining for Discovering Customer Adverse Behavior

This chapter provides a deeper background and motivation of discovering customer adverse behavior, also called adverse selection (AS) behavior, in section 3.1. In section 3.2, I discuss the related work on adverse user identification followed by the proposed framework and description of the method in section 3.3. In section 3.4, I discuss the empirical analysis, which is applied to a real-life insurance dataset to solve the research problem. In section 3.5, I present a comprehensive analysis. Finally, conclusions and future directions are presented in section 3.6.

3.1 Background and Motivation

With the increase in life expectancy (increased from 80.3 to 83.9 years from 2000 to 2020) and increasing pressure on government budgets, life insurance companies play a more significant role in society and provides financial protection to policyholders in need. However, in the life insurance industry, AS behavior of policyholders is typical [22]. It presents a risk to the integrity of the insurance market [38, 61, 208]. Analysis and deep insights into the Australian life insurance market show the existence of adverse activities to gain financial benefits, resulting in loss to insurance companies [38, 155]. For example, a race car driver may acquire a life insurance policy without providing his correct occupational information, even though hiding one's occupation could be criminal.

As another example, a policyholder may receive insurance coverage by providing a residential address that falls within an area with a very low crime rate despite living in an area with a very high crime rate [248]. Insurance companies often lose these misleading practices due to shortfalls in covering the risk. Therefore, detecting AS in the insurance market is necessary to reduce adverse claims and increase business profit and marketing planning [168]. However, the AS behavior analysis of policyholders is challenging, usually involving several factors depending on their preferences and the nature of data, such as the absence of useful private information, the presence of asymmetric information on policyholders, etc. Furthermore, abnormal information exists at the cell level, making it difficult to identify an adverse user.

Insurance companies are often uncertain because of the actions imposed by the unregulated movement of high-risk policyholders [216]. As a result, insurance authorities are keen to understand policyholders' behavior to make appropriate business strategies. Insurance managers (IMs) have started carrying out detective analytics to manage and promote their business efficacy [34]. However, it is very challenging to identify them with some hidden characters related to different data of the insurance policy. There is still a shortage of considerable research regarding detective analytics for the enrichment of the life insurance domain. Existing research has pointed out that traditional techniques are rather time-consuming, taking up to several months, and it is costly to capture comprehensive information on policyholder behaviors. Therefore, it is important for IMs to remain alert for changes, demands, and necessary actions to manage and work with local industries [33, 125, 136]. However, the major challenge for IMs is to keep track of the behavioral patterns of policyholders. Keeping track of policyholder behavior over time is difficult because of its dynamic nature. Behavioral patterns can help IMs make smart decisions that optimize business quality, increase profit, and improve policyholder feedback [94, 192]. Therefore, IMs need to have access to all the critical aspects of the related information, detailing which packages are the best to promote a product, how people prefer different premiums over time, what changes will make a premium more attractive, what actions should be taken to tackle future problems such as the sudden increment of policyholders mental illness claim, a natural disaster and so on.

Advanced data analytics approaches have attracted immense attention from the research community, business decision makers and companies to improve the gain in net profit and have shown considerably better performance via predictive and analytic capabilities [34, 114, 135, 199, 275]. In the insurance industry, while the existing methods explore the hidden behavior of dishonest policyholders, there is still potential to more accurately discover their hidden behaviors. Focusing on these issues, I propose a novel association rule learning-based approach 'ARLAS' to identify the behavior of policyholders. The rationale for taking this approach is as follows: in general, the adverse selection (AS) problem in life insurance does not fit the supervised learning paradigm since there are no labels. Still, life insurers need a method that can identify potential AS behaviors. After consulting with domain experts, I made the assumption that AS behaviors exist but are rare. I recognize that this assumption corresponds to the infrequent patterns in the data set, and such patterns can be extracted using association rule learning reversely, that is, looking at patterns with low confidence but high support. Thus, this approach provides a workaround to make predictions without labels, and the predictions can significantly narrow down the list of suspected AS behaviors to be further verified by insurers.

The main contribution of this study is to propose the first unsupervised learning method to detect AS behaviors in relation to life insurance. This problem can also be viewed as an unsupervised outlier detection problem. Hence, for comparison purposes, I included a few outlier detection techniques such as Local Outlier Factor (LOF), Cluster-Based Local Outlier Factor (CBLOF), One-class SVM, and Isolation Forest (IF) to evaluate the performance of my proposed method. I conducted extensive experiments to study model performance and behavior on one of the largest life insurance data sets ever studied in the literature. The experiment results on the life insurance data of $31,800$ policyholders suggest that association rules can identify AS behavior and assist the insurance authority to reduce loss and guide changes to insurance premium policy for further development management, and planning. The **key contributions** of this work are as follows:

- I present an end-to-end framework to analyze policyholders' adverse behavior that will help the insurance industry reduce the risk of adverse claims.
- I analyze the life insurance user status to identify adverse behavior using ARLAS along with LOF, CBLOF, IF, and One-class SVM.
- To evaluate the performance, I simulated adverse behaviors by randomly flipping the attribute values. I change a random set of 10% (i.e. 318) of the test set records to be adverse-selected and the attributes are reassigned by drawing from the corresponding attribute.

- I analyze 10 years of data on 31,800 policyholders, and create novel association rules that show better performance compared to state-of-the-art methods.

3.2 Preliminary on Adverse Behavior Selection

Machine learning (ML) is mainly used for the prediction and optimisation tasks [30, 31, 172, 306] in life insurance. In this study, I explore the task of detecting adverse behavior (AB). Adverse behavior from policyholders is typical and raises the risk of instability in the insurance market. High-risk policyholders deliberately provide false information to the insurer to escape higher premiums, or to avoid being excluded for eligibility [121, 216]. Existing studies on the AS of the policyholder demonstrate that AB policyholders are better informed about the market likelihood, and use information to select their insurance plans [47, 59, 239]. Additionally, the psychological disorder of the individual can have a deleterious effect on AS behavior. Thus, there is no ambiguity that AS issue has created significant challenges and controversy for insurance industries.

The existing studies by [26, 33, 61] present a clearer view of AS detection. Several studies have highlighted the potential effect of asymmetric information, the proposed methods and key ideas, and have detailed various causes of AS, as shown in Table 3.1. Thus, in the context of the life insurance industry, it has been shown that scrutiny for AS has not extended to the same extent as that for other issues. Cohen and Seigelman [61] reviewed many empirical studies and found evidence of AS in insurance markets, that people in poorer health prefer policies that provide more generous coverage, and policyholders who buy more insurance coverage appear to be riskier [86, 161, 203]. Boodhun and Jayabalan [34] used a supervised machine learning algorithm to assess risk and provide solutions to refine the underwriting process. Boxwala et al. [37] used statistical and machine learning approaches to identify suspicious records. Although engagement classification is related to AS in life insurance markets, there is no analysis of engagement for observing or measuring which individual factors are more likely to cause AS, which individual policyholders are engaged, who are disengaged, and those who are in between [11, 100, 292]. As a result, it is not possible to identify real AS users, and many honest policyholders may suffer. It is worth mentioning that the AS detection method developed in this paper is different from the outlier and anomaly detection methods used in other applications [162, 164, 165, 214, 215, 245, 263, 285, 301, 302] since no explicit labels are provided; instead, I leverage the rule learning technique, which has a long

Table 3.1: Key studies: different methods for adverse-selection detection in the insurance market.

Source	Solution methods	Key Ideas	Purposes
Sengupta and Rooj [239]	Instrument-free semi-parametric copula regression technique	Identification of AS in the healthcare market	To estimate the effect of health insurance status on healthcare utilization
Riddel and Hales [216]	Baseline and control optimism classification model	Risk misperceptions and selection in the insurance market	To investigate the relative influence of baseline and control optimism on selection in an insurance market.
Boodhun and Jayabalan [34]	Supervised learning algorithms	Risk prediction in the life insurance industry	To classify the risk level by applying a predictive model.
Keane and Stavrunova [136]	Smooth Mixture of Tobits	Analyze AS and moral hazard in a unified economical framework	To estimate AS and moral hazard effects jointly in the Medigap market.
Song et al. [246]	Machine learning methods	Assess financial fraud risk	To identify the risks associated with financial fraud, and help to reduce enterprise risks.
Meyer et al. [182]	Data mining classification techniques	Improve the dynamic decision strategies.	To discover treatment strategies by predicting and eliminating treatment failures.
Boxwala et al. [37]	Statistical and machine-learning approach	Identify suspicious records	To help privacy officers detect suspicious access to EHRs.
McCarthy and Mitchell [181]	A over E	Adverse selection in life insurance and annuities	To assess the extent to which life insurers can hedge mortality exposure by writing both life insurance and annuities.
He [100]	Conditional correlation approach	Find the relationship between a high-risk and low-risk person	To examine the presence of AS in the life insurance market.

history but still shines in recent works [318].

From the above review of the existing studies, it is clear that most of the methods focus on limited aspects and were limited in their performance and capability. For instance, [69, 88, 181, 246] provide evidence for AS in insurance markets but they used limited information. However, there are many aspects associated with demographic and socio-economic information such as age, postcode, occupation, and gender, which have made insufficient research concern for AS purposes [14]. They go on the AS hypothesis test using statistical models, which could cause bias in the results of the estimation. Yet importantly, there have been some attempts to used data mining and machine learning to analyze and propose solutions using policyholder data within the life insurance industry [34, 113, 122]. To analyze and describe potential predictive factors, they use straightforward regression models. However, the predictive performance of the existing techniques is rather low. While an increasing body of research combines insurance data with machine learning techniques to observe policyholder behavior, it is challenging to do this with sensitive policyholder data and there is scope to apply advanced machine learning and data analytics techniques.

In summary, the existing research on AS behavior analysis is limited. Very few studies have considered advanced data analytics techniques to meet the practical requirement of the insurance industry. Furthermore, a very limited number of datasets have been used in literature. To deal with the aforementioned challenges, in this work, I analyzed 10 years of data on 31,800 policyholders, and propose the first unsupervised learning method for detecting AS behaviors in the life insurance industry.

3.3 Methodology

In this section, I first describe the details of data collection and processing. I then present the proposed framework and method for the detection of AB in the life insurance industry.

3.3.1 Data Collection and Processing

I use two types of datasets, namely (i) questionnaire based behavioral data, and (ii) demographic data. I collect the data from one of the most popular insurance companies in Australia, where users are required to answer various questions. The behavioral

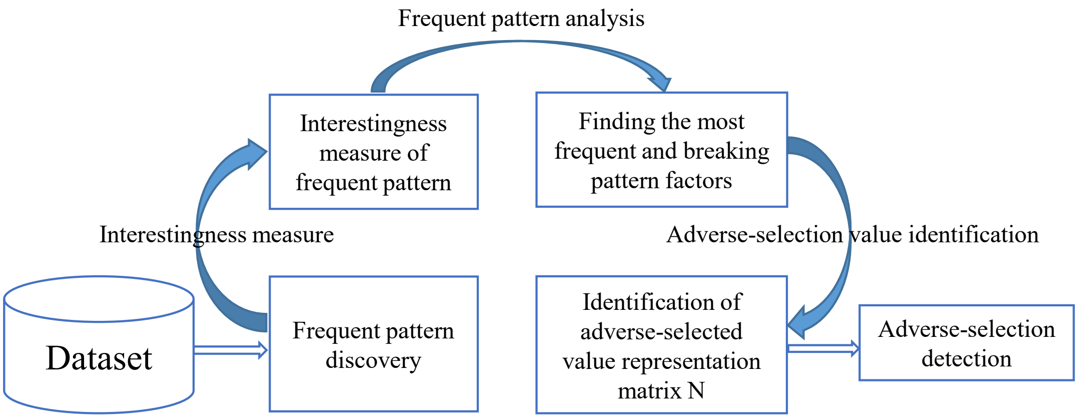

Frequent pattern analysis
Interestingness measure of frequent pattern
Finding the most frequent and breaking pattern factors
Interestingness measure
Adverse-selection value identification
Dataset
Frequent pattern discovery
Identification of adverse-selected value representation matrix N
Adverse-selection detection

Figure 3.1: The structure of the *ARLAS* framework.

dataset contains 31,870 data records related to one of the insurance applicants and includes 834 columns, each pertaining to a yes or no question. On the other hand, the demographic dataset contains information on the policyholders' ID, gender, postcode, age, and occupation. When the dataset was ready, I started processing the data. Before any data analysis process can begin, the dataset requires cleaning and pre-processing to remove ambiguity. Any ambiguity or confusion in the dataset can lead to an incorrect analysis. Therefore, I wrote Python scripts to start the data cleaning process and cleaned the dataset. Finally, I resolved missing and invalid data, and all data was subjected to a quality test.

3.3.2 Proposed Model

In this section, I present a model *'ARLAS'* to detect the AS behavior of users in the life insurance market as shown in Figure 3.1. The approach is similar to the method proposed by [94], which has been applied to smart home data for behavior monitoring and abnormality detection. I use a frequent pattern mining algorithm which has the ability to locate repeating relationships between unique items in a data set and represent them in the form of association rules. To analyze insurance data to identify the AS behavior of policyholders, I carry out the following steps as listed below:

3.3.2.1 Frequent Pattern Discovery

The apriori-based frequent itemset algorithm is used to mine frequent itemsets to generate patterns [6]. It uses an iterative level-wise search technique to discover the $(k+1)$ item sets from k-item sets, for example, a sample of the questionnaire database that comprises the various questions answered by different users. First, it scans the database to identify all the frequent itemsets by counting each of them and capturing those which satisfy the minimum support threshold. The identification of each frequent itemset set requires scanning the entire database until no more frequent k-question sets are identified.

3.3.2.2 The Interestingness Measure of the Frequent Pattern

To illustrate, I assume that the formal description of a frequent pattern is as follows:

$$(A \rightarrow B) \tag{3.1}$$

In this description, $A = \{a_1, a_2, a_3, \ldots, a_n\} \in I$ and $B = \{b_1, b_2, b_3, \ldots b_n\} \in I$. I show itemsets and $A \cap B = \phi$. The patterns should meet a certain support threshold s. Therefore, according to [129], the standard five measures are used to characterize the frequent patterns.

Support: For a transaction set D, the support of an itemset X is given by

$$\text{supp}(X) = \frac{|t \in D; X \subseteq t|}{|D|} \tag{3.2}$$

Confidence: Confidence is the conditional probability of subsequent occurrence as a result of the previous data. The rule $(A \rightarrow B)$ has confidence c in the transaction set D, where c is the percentage of transactions in D containing A that also contain B, i.e.,

$$\text{conf}(A \rightarrow B) = \frac{\text{supp}(A \cup B)}{\text{supp}(A)} \tag{3.3}$$

Lift: Lift refers to the ratio of the occurrence probability of B under condition A to that without considering condition A, which reflects the relationship between A and B. The interest of the rule $(A \rightarrow B)$, also known as lift, is:

$$\text{lift}(A \rightarrow B) = \frac{\text{supp}(A \cup B)}{\text{supp}(A) \times \text{supp}(B)} \tag{3.4}$$

Leverage : A new interestingness measurement method of ARL is based on the description of the defects of the traditional interestingness measurement method. The leverage of the rule $(A \rightarrow B)$ is defined as:

$$\text{leverage}(A \rightarrow B) = \text{supp}(A \cup B) - \text{supp}(A) \times \text{supp}(B) \tag{3.5}$$

Conviction: The conviction of $(A \rightarrow B)$ is defined as:

$$\text{conviction}(A \rightarrow B) = \frac{1 - \text{supp}(B)}{1 - \text{conf}(A \rightarrow B)} \tag{3.6}$$

3.3.2.3 Finding the Most Frequent Patterns and Frequent Pattern-Breaking Factors

A dataset contains many factors used to create distinct patterns. However, not all factors can create patterns all the time. The factors are more informative when they plays an important role in creating the pattern. Therefore, the factors that are used to create a pattern frequently are the correct factors. In contrast, when it breaks, the most frequent patterns are the adverse-selected factors (ASF). For example, suppose $D = \{t_1, t_2, t_3, \ldots, t_n\}$ is a database containing a set of n items $I = \{i_1, i_2, i_3, \ldots, i_n\}$. An itemset X is a non-empty subset of I. Given a minimum support threshold, minisupp, find all itemsets when they break the rules with supports greater or equal to minisupp.

I created a list of the frequent patterns (see 3.4.2). I extracted the most frequent pattern through a user-specified minisupp threshold value. In this step, the user-specified support threshold is set to 0.015. I test the support threshold with different sizes ranging between 0.001 and 0.030 where minisupp 0.001 extracts many patterns but affects the execution time and minisupp 0.030 extracts very few patterns. Therefore, by setting the minisupp threshold to 0.015, I decrease the execution time and obtain a reasonable number of patterns. Additionally, to decrease execution time, I omit patterns with lengths greater than four. I then determine how often the factors used to create the most frequent pattern fail. The initial assumption of the breaking frequent patterns (BFP) is that when the factors used to create the most frequent pattern fail, I identify them as the breaking frequent patterns of factors. The set of all breaking frequent patterns is denoted by BFP(D, minisupp), i.e.,

$$\mathrm{BFP}(D, \mathrm{minisupp}) = X \not\subseteq I | \mathrm{supp}(X) \geq \mathrm{minisupp}. \tag{3.7}$$

For each transaction t, the frequent pattern ASF of t is defined as:

$$\mathrm{ASF}(t) = \frac{\sum_{X \subseteq U, X \in \mathrm{BFP}(D, \mathrm{minisupp})} \mathrm{supp}(X)}{||\mathrm{BFP}(D, \mathrm{minisupp})||} \tag{3.8}$$

The interpretation of equation 3.8 is: if I contains more breaking frequent patterns, its *ASF* value will be large, which shows that it is more likely to be an AS factor. In contrast, the factors with small *ASF* values are unlikely to be an AS factor. Obviously, the *ASF* value is between 0 and 1.

3.3.2.4 Analysis of Adverse Selection Detection

In this step, AS factors and high-risk policyholders are detected. I first construct a frequent pattern value matrix and further transfer it to the AS value matrix by breaking rules that are interrupted to generate the frequent pattern. I define the AS value matrix N as follows.

$$N = \begin{pmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,n} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m,1} & a_{m,2} & \cdots & a_{m,n} \end{pmatrix} \tag{3.9}$$

where $a_{m,n} = ASF$ value which is mentioned in 3.8.

I extracted the breaking pattern value of the factors and constructed the $n \times m$ matrix where the row shows the user and the columns show the AS factors. But the different factors in the matrix have different values. I then transferred the values of the factor to a common scale by normalization. I did this to change the values of numeric columns to a common scale, without distorting differences in the ranges of values using the following equation:

$$X_{new} = \frac{X - X_{min}}{X_{max} - X_{min}} \tag{3.10}$$

where X_{new} is a set of the re-scaled values present in X that will now be $0 \leq X_{new} \leq 1$, X_{min} is the minimum values in X and X_{max} is the maximum values in X. The greater the breaking frequent pattern values of the factor, the higher the probability of it being an AS factor. Therefore, when a user has many AS factor values, they are more likely to engage in AS.

3.3.2.5 Computation Complexity

The proposed association rule learning-based approach is based on frequent pattern mining which is an NP-hard problem [297]. Thus, the complexity of the proposed framework can be determined by the frequent itemset mining algorithm. Therefore, the complexity of the proposed method is $O(nN^2)$, where n represents the data records and N represent the number of items.

3.4 Experiment

In this section, the data description, interesting pattern list construction, the results and the analysis of my proposed model for solving the AS problem are addressed, respectively.

3.4.1 Data Description

I utilized the following datasets for the empirical works in this study. I applied these datasets to provide a broader and more comprehensive analysis of AS behavioral modeling in the insurance industry. I processed all the datasets to remove any personal identification, including anonymizing names and personal contact details used in the datasets.

Questionnaire Dataset: I acquire the dataset from a screening questionnaire provided by a local Australian life insurance company. The questionnaire was large and detailed, conprising data on $31,800$ users and each user answered 834 questions ranging from personal details, lifestyle, and family history to occupational details. The data is binary data, where if the applicant answered 'yes' to the question, the cell contained a '1', and if the user answered 'no', it contained a '0'. For example, if a user drinks alcohol, the 'alcohol' attribute must contain a '1' in the dataset, and if the user does not drink alcohol, the attribute must contain a '0'.

Demographic Datasets: As shown in Table 3.2, there are five different variables in the demographic dataset such as policyholder life insurance ID, gender, age, occupation, and policyholder postcode for the $31,800$ policyholders. The 'Gender' attribute is denoted as either 'M' or 'F' for 'Male' and 'Female' respectively. The 'Postcode' attribute contains the Australian postcode of the applicant's residence. The 'Age' attribute contains the age of the applicant in whole years, where the youngest applicant is 3 years old and the oldest is 78 years old. The 'Occupation' attribute contains 18 different categories. Examples of these include 'T-Trades', 'S-Supervisor of Trades', 'R-Special Risk', 'Q-Qualified Professional', 'OR-Ordinary Rates', 'L-Light Trades', 'H-Heavy Trades', 'F-Financial Professional', 'D-Medical/ Dental', 'I-Indoor Sedentary', and 'C-Community Professional'. As part of the demographic information analysis, I use the Socio Economic Indexes for Areas (SEIFA) data set to rank areas in Australia by relative socio-economic advantage and disadvantage [207].

Table 3.2: Characteristics of demographic dataset.

Variables	Category
Policyholder life insurance ID	Number
Gender	Male Female
Age	(1-15) (16-30) (31-45) (46-60) (61-75)
Occupation	18 different categories
Postcode	Number
Socio Economic Indexes for Areas	Index of Relative Socio-economic Advantage and Disadvantage (IRSAD) Index of Relative Socio-economic Disadvantage (IRSD) Index of Economic Resources (IER Index of Education and Occupation (IEO)

3.4.2 Interesting Pattern List Construction

In this section, I apply the proposed frequent itemset mining algorithm to the collected data. In the proposed framework, I exploit the user-defined minimum support threshold minisupp to imply the ‘minimum frequency’ for ‘ARLAS’ model construction and to determine whether there are valid relationships in the provided data. The support value dictates how frequently a particular itemset appears within a dataset where an itemset with higher support justifies greater commonality or popularity. Therefore, I explore the effect of setting various support thresholds, ranging between 0.001 and 0.030, with respect to the number of features identified as candidates.

As shown in Figure 3.2, the algorithm identifies $66,000$ patterns with minisupp 0.001, which is all the patterns in the stemmed list. The number of patterns gradually decreases to $4,000$ when minisupp is set to 0.015, then continues to decrease slightly with an increase in support thresholds. When minisupp is 0.030, only 147 patterns are returned. Notably, pattern generation is automated; thus, users may consider the output to choose their pieces of interest. This method is suitable in the process because the pattern number is mostly small. Hence, this condition provides a better extensive and consistent list of patterns to be formed.

Table 3.3: Different interestingness measure of the different patterns.

Patterns	Supp.	Conf.	Lift	Lev.	Conv.
1. Screening off work 7days minor → Screening off work 7days minor recovered	0.001	0.98	68.41	0.001	50.76
2. Screening Musculo Skeletal Back, Screening Skin Lesion → Screening Consult Test Prescription	0.002	0.98	17.30	0.002	77.79
3. Screening Neurological, Screening Musculo Skeletal Joint → Screening Consult Test Prescription	0.001	0.98	17.27	0.009	66.94
4. Screening Respiratory Asthma, Screening Sensory Eyes → Screening Consult Test Prescription	0.001	0.98	17.32	0.001	85.80
5. High BP Medication, Screening Sensory → Screening Cardio High BP	0.003	0.98	40.12	0.003	49.55
6. Screening consult test , High BP medication → Screening cardio high BP	0.002	0.99	40.54	0.002	147.30
7. Screening consult test prescription, High BP medication → Screening cardio high BP	0.002	0.99	40.54	0.002	145.34
8. Joints area arm, Mental health specialist referral, Screening consult test prescription → Mental health medication	0.001	0.98	20.21	0.001	82.75

Several other measurement methods such as ‘confidence’, ‘lift’, ‘leverage’ and ‘conviction’ are applied where various factors of interest to users are also identified. These detailed aspects include ‘family history’, ‘life style’, ‘insurance history’, ‘employment information’, ‘medical history’, ‘health and risk factors’, and ‘socio-economic factors’ such as ‘remote area’, ‘family type’, ‘gender’, and ‘age’ which are significant to insurance managers. Table 3.3 shows some interesting patterns’. These results are impressive because these terms are related to policyholders’ behavior information.

3.5 Result and Discussion

3.5.1 Result Analysis

I propose a novel ARL based method to detect the AB of policyholders in the context of life insurance. I compare the findings with other established unsupervised learning methods which are used for similar analysis such as LOF, CBLOF, IF, One-class SVM. By

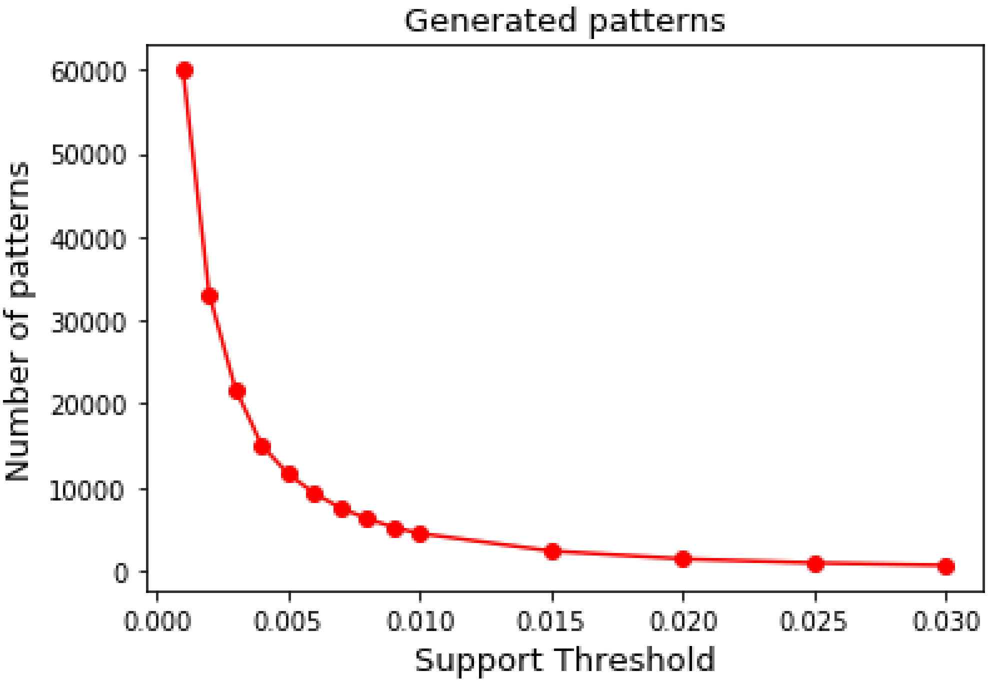


Figure 3.2: Identifying different rules with different support threshold.

comparing and contrasting the results, I determine whether it produces similar results. However, without label data I cannot evaluate the performance of the models using simple measures such as accuracy, ROC, AUC, etc. Therefore, in this research, I have selected unsupervised learning measures, such as the silhouette (SI) score to evaluate the performance of the baseline models. To evaluate the performance of the proposed method, I create synthetic AS results by randomly flipping the attribute (RFA) values. I begin by partitioning the dataset into training and testing sets. I sample 10% of the data for the test set, and the remaining 90% is chosen as the training set. The original dataset used to generate these results has 31,800 records, so after partitioning process the training set has 28,620 records and the test set has 3,180. I change a randomly sampled 10% (i.e. 318) of the test set records to be AS. For each record that is modified, I choose a random set of up to l attribute. The values for these attributes are reassigned by drawing from the corresponding attribute marginally. The higher the value of l, the greater the degree of AS. However, these scores only evaluate the performance of the models according to the inter- and intra-distance measures and they cannot evaluate the models according to business requirements, thus it is the best option for evaluating the manual validation by insurance professionals. The results and findings of these methods are obtained and compared in Table 3.4 and I provide further details as follows.

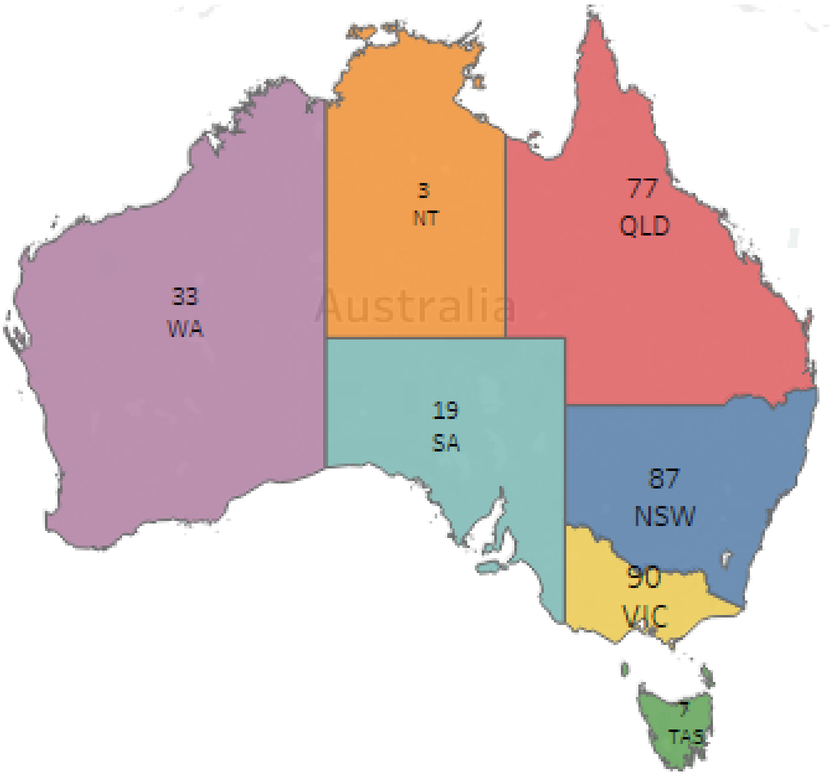


Figure 3.3: Number of adverse policyholders per state.

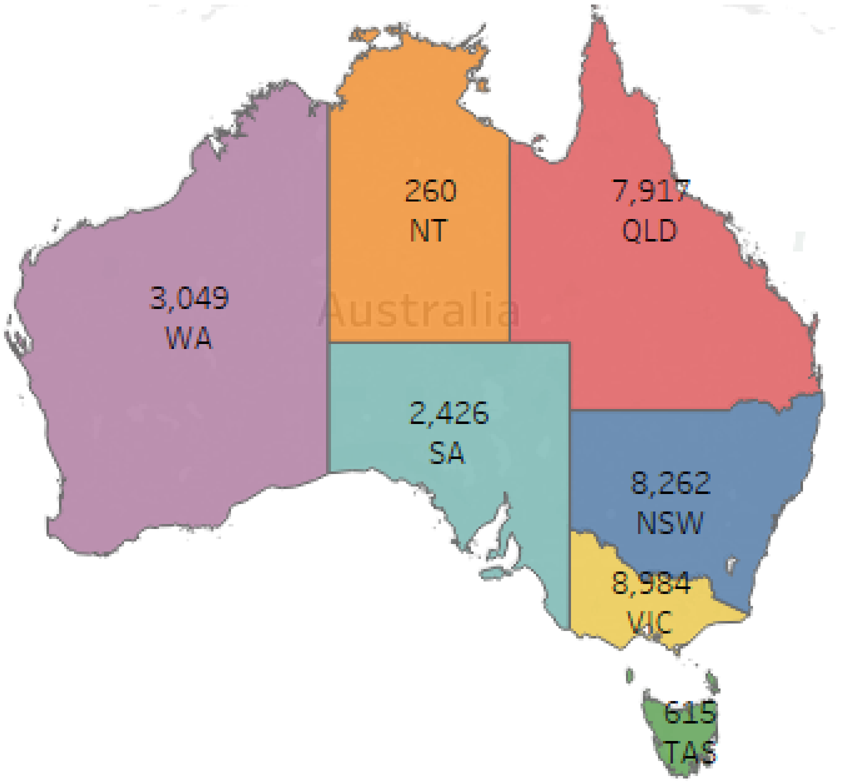


Figure 3.4: Number of policyholders per state.

Table 3.4: The result of the experiment.

	LOF	CBLOF	Isolation Forest	One-class SVM	**ARLAS**
SI	0.50	0.49	0.58	0.58	–
RFA	–	–	–	–	**0.63**
Number of generated patterns	–	–	–	–	4000
Number of clusters	–	15	–	–	–
Total no. of policyholder	31,800	31,800	31,800	31,800	31,800
Number of adverse policyholders	296	301	307	308	319

I provide a detailed analysis based on applying the rules to the whole data set to analyze the AS patterns, as shown in Section 3.5.1. I do not use the synthetic AS data in the rest of the paper. Therefore, through the extensive analysis, I visualize the distribution of AS policyholders. Using the list of AS policyholders derived from ARLAS, I look at the distribution of locations of individual users. The list derived from the approach gives me proportions of 29.15% (90 users) of risky users from Victoria, 26.96% (87 users) from New South Wales, 23.51% (77 users) from Queensland, 10.66% (33 users) from Western Australia, 5.02% (19 users) from South Australia and the remaining 4.7% are spread evenly between the Australian Capital Territory, the Northern Territory, and Tasmania. From this information, it is clear that these figures are correlated with state and territory populations, except for Victoria, which produce a higher proportion of risky users relative to its population of around 6.5 million compared to roughly 8 million in New South Wales. Looking further within New South Wales, I divide the risky users based on more precise locations and discover that a large portion of risky users come from the inner west and eastern suburbs, this being 23.3% and 19.8% respectively.

Figure 3.5 highlights the different occupations of these AS users, and the number of people within each occupation pertaining to the results derived from the proposed approach. I found that the highest percentage of applicants worked in indoor sedentary occupations with 21.9% of AB users (70 individuals) making up this section. An indoor sedentary occupation is defined as any job where the employee spends most of their time sitting down. This covers most desk jobs, and jobs at call centers, professional drivers such as bus drivers, taxi drivers, truck drivers, train conductors and pilots, software developers, accountants, and designers. An article released by [103] provided evidence that sitting down for extended periods of time has been linked to a variety of health

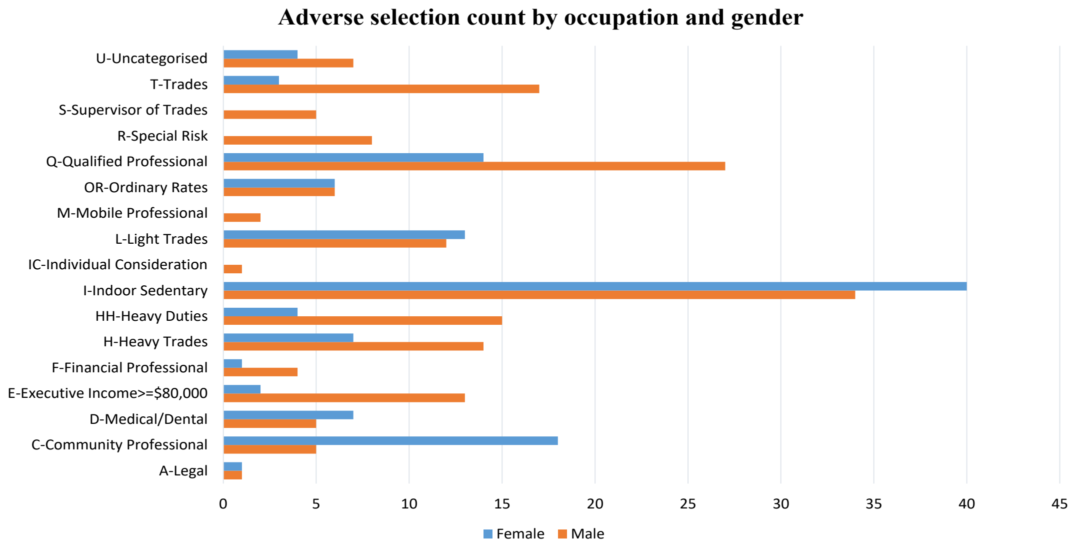


Figure 3.5: Number of adverse policyholder by occupation and gender.

risks and diseases such as “obesity, diabetes, hypertension, and heart disease.” With all these health risks linked to indoor sedentary occupations, life insurance companies consequently charge higher premiums for this type of occupation while additionally charging more if applicants encounter such health risks.

Next, I categorized the AB users by age. Figure 3.6 and 3.7 visualizes the AB users within the age ranges to see which age range contains the most problematic users. I discovered that a considerably large portion of applicants fall within the $31-40$ range, with 130 people (40.75%) making up this category. Additionally, 63.23% of AB users (202 people) are aged between 31 and 50. Studies have shown that both males and females in this age bracket have increased chances of diseases such as heart disease, obesity, cancer, and diabetes, which consequently increases insurance premiums if diagnosed. Because of health issues and disease, these being the main contributors to increased premiums, those who wish to avoid such an expense are more inclined to deny being at risk of these diseases and also have the incentive to lie in an application regarding such health issues. This results in the aforementioned consequences that affect not only the insurance company but also other applicants who are forced to pay more to account for the claims made by AB users. Another interesting finding that I made after the visualization process was that 23 AB users (7.2%) were in the < 20 age category. Young people in this bracket rarely have health issues that warrant life insurance. However, if these are exceptional cases, it would make sense why the detection methods detected

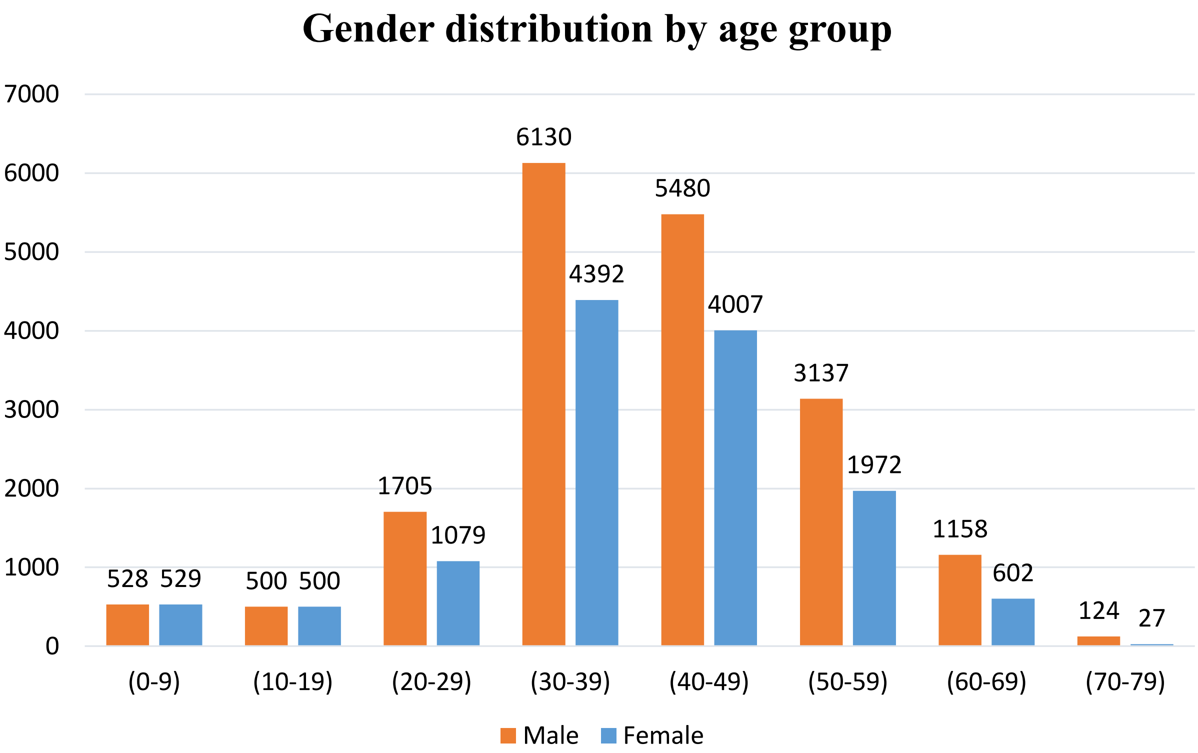


Figure 3.6: Number of policyholders by age group.

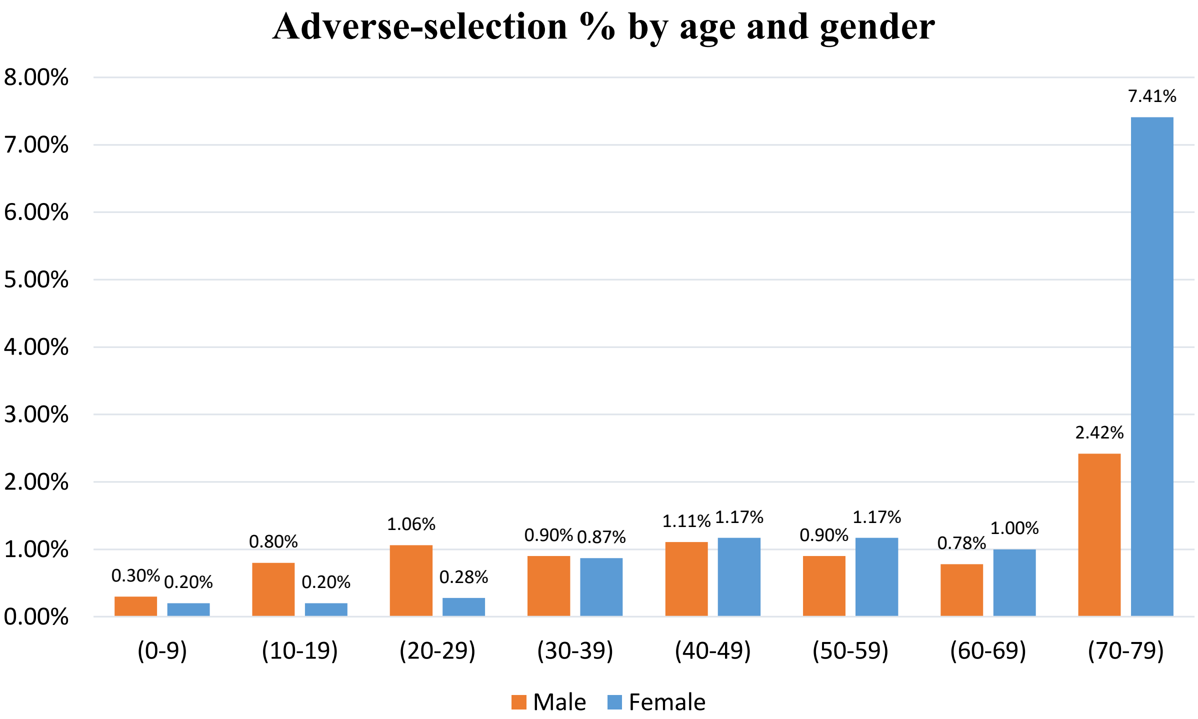


Figure 3.7: Number of adverse policyholders by age group.

them as being adverse answers.

3.5.2 Discussion and Implications

In this section, I provide a brief discussion and the implications of AS detection. To identify the AB of policyholders accurately, I consider both the questionnaire-based behavioral and demographic data of Australian life insurance policyholders. Earlier studies on Australian life insurance mainly focused on statistical approaches. However, in this work, I use the ARL-based approach to explore the AB in depth to have a better understanding of what the data represents, the behaviour of the adverse policyholder, how the adverse policyholder differs from real users, etc. Through my research work, I found that the life insurance industry is at risk in Australia. To manage insurance data, the insurance authority needs to have a comprehensive understanding of normal and risky policyholder information and then be able to identify AB behavior. Therefore, I describe a model to obtain the details of policyholders who help to identify AB.

The analysis of the demographic information in Figure 3.3 and 3.4 suggests that IMs should pay more attention to NSW and VIC where policyholders stand to receive considerable benefit. Brisbane is also a high-potential area, where high-risk policyholders spend a long time. Therefore, IMs can investigate and develop business strategies among policyholders when they buy their insurance policy and reduce insurance loss.

The behavior analysis of policyholders provides an example of how different professional information can be extracted and analyzed for valuable insights. Prior studies often focused on the significant factors [96, 161]. However, less significant factors should also be given attention because they can generate a substantial profit for high price ranges. Therefore, the occupation distribution in Figure 3.5 helps IMs to realize the fact that premiums may rise significantly based on the profession of a policyholder.

The analysis of the age distribution of males and females shown in Figure 3.7 is necessary for IMs in designing appropriate policy packages for the future. It shows that the average age of the adverse policyholder derived from the proposed method is 41 years old. The difference in the ages of males and females is higher in the age bracket (70 – 79) but in the age bracket (40 – 50), both are almost equal. Female policyholders are more adverse than males; they have less income but higher consumption expenditures than a

Table 3.5: Advantages and limitations ML/DL techniques.

Techniques	Advantages	limitations	References
Logistic Regression (LR)	Simple to implement and more accessible to interpret the output.	Not useful for complex analysis.	Connelly et al. [64]
K-Means Clustering	Computationally faster than hierarchical clustering.	Difficult to predict K-Value with the global cluster.	Mai et al. [177]
Long Short Term Memory (LSTM)	Reliably transmits essential information into the future in multiple iterations.	Difficult to train because they require memory-bandwidth-bound computation, which is the limitation of neural network solutions.	Demir et al. [73]
Convolutional Neural Network (CNN)	CNN facilitates autonomous classification of glaucoma based on complex features derived from thousands of available fundus images with specificity and sensitivity ranges between 85 to 95%.	Overfitting, explosive gradients, and class imbalance are the most common problems encountered using CNN to train the model.	Gheisari et al. [93]
Graph Convolutional Networks (GCN)	Scalable solution for large-scale graphs and high accuracy with a low label rate.	Complex data models.	Tian et al. [267]

male policyholder.

During my research, I encountered several limitations such as the availability of the required datasets, the implementation of some other more accurate machine learning and deep learning models such as logistic regression, K-Means clustering, long short term memory (LSTM), convolutional neural network (CNN), and graph convolutional networks (GCN) as shown in Table 3.5. There is no standard label data available, thus there is a strong need for a dataset for supervised learning aided by expert knowledge. The results will be more accurate with a huge dataset where all the policyholders information is confirmed. The behavior preference is more applicable and practical when the dataset is huge, which could be one of the limitations of the research that I found during the analysis phase. On the other hand, I only focused on the analysis of user behavior in Australia before COVID-19 since there was not much user information due to lockdown in Australia and many other countries. Additionally, the proposed approach has some limitations. First, Apriori-based frequent itemset mining generates large candidate sets and repeatedly scans the database, which requires a lot of run time and memory. Second, in frequent itemset mining, the order of items in the itemsets is unimportant. However, there are some situations in which the order of items inside the item is important. Third, if a pattern is frequent, its sub-patterns are also frequent. However, there are some cases where patterns and sub-pattern are not the same.

I make the following **key observations**

- The extensive study of 10 years of data of $31,800$ policyholders showed that, for the age range $31-50$, the number of adverse-selected female policyholders is considerably higher than male policyholders, thus IMs should pay more attention to female policyholders in NSW and VIC.
- This study suggests that, premiums may rise significantly based on the profession of the policyholders.
- I note that the average age of the adverse-selected policyholders is between 35 and 45 years old. The risk of adverse claims for this population can be reduced by considering other factors.
- For a larger dataset, behavior preference could be used to improve the performance of AS.

3.6 Summary

Understanding the behavior of policyholders is necessary to reduce AS, increase business profit, and improve marketing planning. Therefore, I proposed the first unsupervised learning method for detecting AS behaviors in life insurance. I conducted extensive experiments to study the proposed method's performance and behavior on one of the literature's most significant life insurance data set ever studied. A comparison and evaluation of real-world insurance data showed that the proposed approach showed considerable gains in performance by identifying 319 adverse cases compared to 296, 301, 307, and 308 using LOF, CBLOF, IF, and One-class SVM. This research also lays out a fundamental framework and structure to support further research on such topics. Being able to recognize a future trend in the insurance industry would help IMs in the decision-making process.

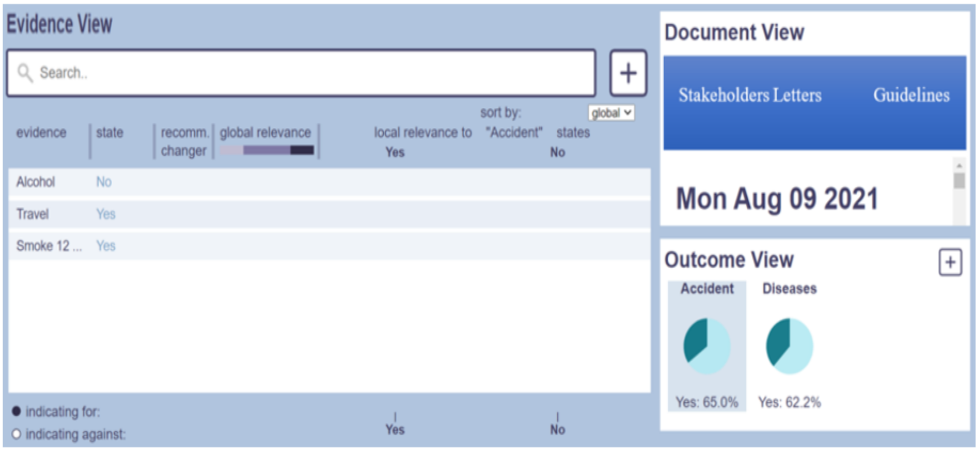
CHAPTER 4

Visual Analytics for Exploring Customer Behavior

In this chapter, I provide the background and motivation of visual analytics for understanding customer behavior analysis in Section 4.1. In Section 4.2, I discuss preliminary work on customer behavior exploration, techniques, and behavioral data visualization followed by the detail methodological discussion in Section 4.3. The description of the visual analytic solution (*ExVis, MHIVis and DiaVis*) are presented in Section 4.4. I illustrate user study of the proposed visual analytics system (VAS) to assess and discuss its capacity to inform the relevant variables for exploring customer behavior in Section 4.5. Finally, conclusions and future directions are provided in Section 4.6.

4.1 Background and Motivation

Exploring customer behavior with multi-dimensional and temporal data is necessary for any competitive and global business to provide exciting insights and improve business strategies. For example, the financial industry, such as the life insurance industry, is becoming complex day by day, where effective communications between risk experts and decision-makers play a vital role [67, 120, 197]. While existing researchers have applied various data analytics approaches to understand and analyze customer behaviors, they often failed to allow the analysts, including business management, product marketing, development, decision-making, etc. Visualization is one of the efficient ways to support



evidence	state
Alcohol	No
Travel	Yes
Smoke 12 ...	Yes

Figure 4.1: *ExVis*: exploring multiple views to explain visual decision support system.

such effective risk-related communication [112, 124]. Nowadays, many types of business strategies, mapping approaches, and visual metaphors are employed in various business purposes [85]. However, visual analytics for exploring customer behavior is still rare in the business community because the risk is both challenging to visualize and hard to manage.

A fundamental business aspect of the business market is assessing the risk connected with each individual [160]. The financial industry, such as the life insurance industry, is becoming complex, where effective communications between risk experts and decision-makers play a vital role [67, 120, 197]. For example, policyholders with mental health illness (MHI) cause a financial imbalance and affect their professional and personal lives [120, 121]. The existing studies show that the qualitative and quantitative aspects have been dominated to analyze and manage the risks. Tables, charts, maps, and formulas are standard tools in daily communication between experts and managers regarding financial risks and how to handle them [85]. However, it is challenging for decision-makers to identify and understand the most significant risks and implement adequate remedies when considering policyholders' negative experiences seeking insurance benefits. Therefore, there is no doubt that insurance companies make a loss, where an interactive solution is required to justify the claiming insurance benefits with proper explanation.

Mental Health Illness (MHI) is a significant issue in the Australian landscape, with an ever-increasing number of cases [65]. Almost one in five Australians within the age

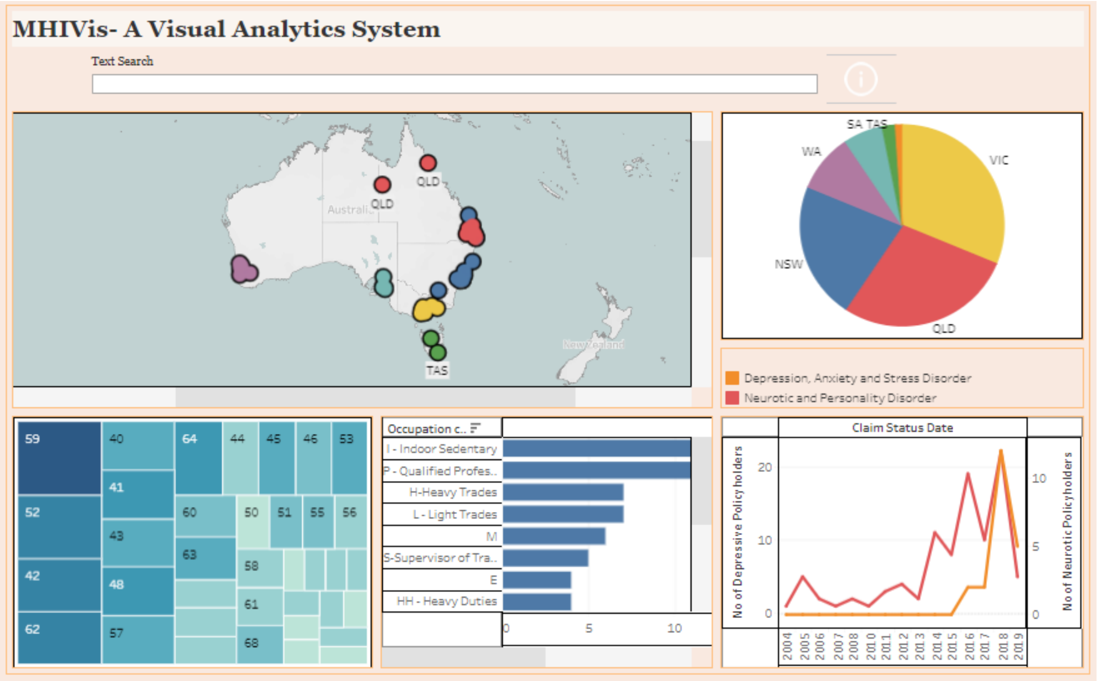
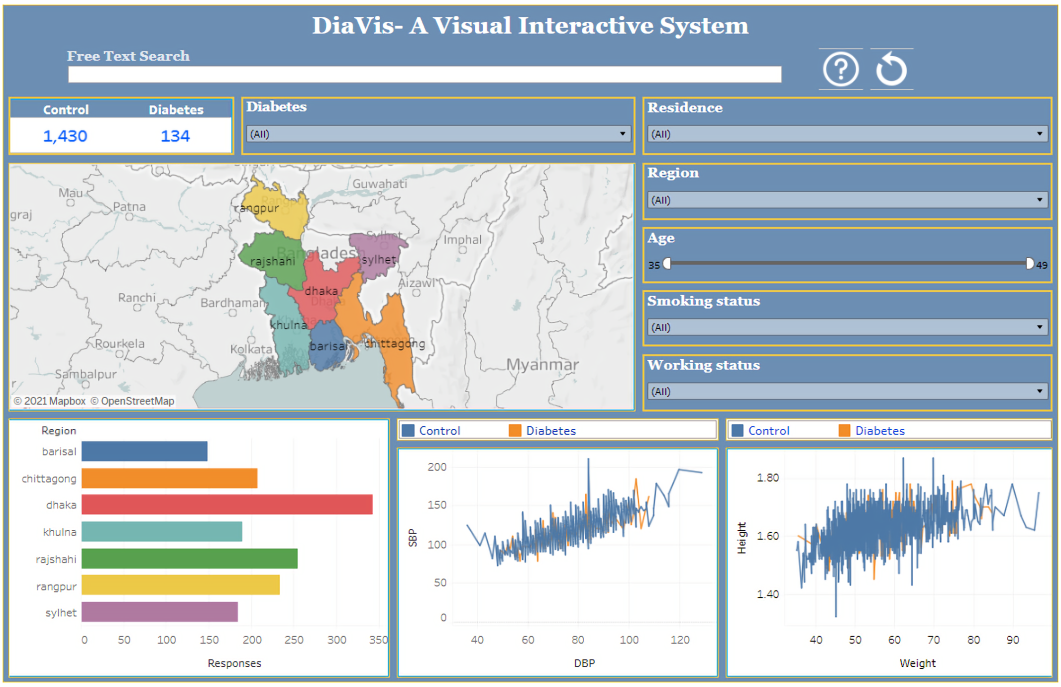


Figure 4.2: *MHIVis*: a visual interactive system for exploring mental health illness.

range 16 to 85 have MHI [21]. MHI can have a deleterious effect on the physical and psychological health and the productivity of the work of an individual [21]. Besides, it is the main reason for claiming injury compensation in Australia [222]. So, there is no doubt that mental illness has brought enormous challenges for the insurance industry. Assessing the MHI associated with any insurance applicant is a core business characteristic of the insurance market [120]. People with MHI bring forth a financial imbalance, superfluous working spirit, alongside affecting their professional and personal life [123]. Therefore, to understand the MHI of Australian people, I consider a local life insurance policyholder's MHI. In the life insurance industry, it has been shown that policyholders with MHI have three times higher rates to claim for benefit compared with individuals who do not have a MHI [222, 270]. This kind of enduring behavior is referred to as underwriting sometimes and a key to ensuring the economic viability of the insurance industry, which causes a significant impact on the economy. When considering the negative experiences of policyholders in applying for claiming insurance benefits, it is necessary to consider the personal information and risk assessments of MHI policyholders in the community [130]. However, sometimes stakeholders are confused to deal with the complexities of MHI and observe which individuals are engaged, which are disengaged, and those who are in between. Therefore, there is an urgent need to identify a group of professional people in a specific age range and their living areas (dangerous) in Australia



DiaVis- A Visual Interactive System

Free Text Search

Control 1,430
Diabetes 134

Diabetes: (All)

Residence: (All)

Region: (All)

Age: 35 – 49

Smoking status: (All)

Working status: (All)

Figure 4.3: *DiaVis*: a visual interactive system for exploring diabetes disease.

to help with special considerations and profitability of those experiencing mental illness.

Diabetes mellitus (DM) is a hormonal and metabolic disorder in which the body can not produce enough insulin and increase blood sugar level abnormally high [191]. As a result, diabetes damages the nerves, raising the risk of chronic kidney disease, stroke, heart attack, eyesight loss, and so on. Diabetes mellitus is a syndrome that is now recognized and classed as a disease defined by signs and symptoms of chronic hyperglycemia. The number of diabetes patients is increasing day by day, where 1.6 million people were died due to diabetes [28]. According to the statistics, it has been stated that about 122 million people affected by diabetes in 1980, with the rate increasing to 422 million in 2014 [62]. Furthermore, the estimation will be struck to 642 million approximately in 2040 [319]. Over the last 20 years, the number of patients significantly affected by type 2 diabetes, approximately 90%. This increasing rate is very much alarming for the future. In response, there is a growing need for identifying diabetes disease and the significant factors that have a substantial impact on diabetes. Moreover, delays in diagnosis are an important contributory factor to poor control and risk of complications.

Visual interactive systems (VIS) are widely used for different purposes such as recommendation systems, mental health analysis, disease analysis, sentiment analysis, etc.

[122, 137, 313]. For example, visualizing healthcare data [289], Hadoop based analysis and visualization of diabetes [29], were emerged to provide a visual overview of diabetic disease analysis. Additionally, Varga [272] demonstrated how visual analytics (VA) could help financial services thrive. Islam et al. [121] designed a new VIS named *MHIVis* to provide a visual summary of mental health disorder data. However, their contribution focused on structured data, whereas the implementation provides an interactive and effective dashboard that flourish more insights regarding diabetes disease. Additionally, some VIS still incomprehensible and less appealing in many situations when it comes to decision-making [189, 190]. As a result, the concerns of presenting the data and why a particular outcome is recommended remain unresolved [99, 317].

To address the aforementioned challenges, first, I design an interactive explainable visual decision support system named *ExVis* for risk management that combines a Bayesian network (BN) model with visual recommendations. Figure 4.1 illustrates how *ExVis* enables IMs to assess policyholder history to recommend future claim records. After that, I design a new visualization system named *MHIVis*, visual analytics for exploring mental health illness, which allows the stakeholders to type different text queries such as 'Find top 3 MHI states' in a text search box within a dashboard. Figure 4.2 illustrates how the system helps stakeholders understand policyholders' MHI. The dashboard evaluates 10 years of data from 31,800 policyholders to provide a coordinated view of accidents and diseases across Australia. Besides, I designed an interactive visual system (Vis) for exploring diabetes disease named *DiaVis* to address the limitations mentioned above as shown in Figure 4.3. The goal of using the visualization technique is to give a more comprehensive solution, to assist end-users in analyzing diabetes conditions, as well as to identify the key factors which are significantly involved with diabetes disease. With that aim, first, I employed a case study model to understand the factors that affect the diabetes disease as shown in Figure 4.8. By observing the increasing rate of diabetes patients and previous research of diabetes disease, I were motivated to propose *DiaVis* that can represent the significant diabetes factors. *DiaVis* represents multiple relations with diabetes such as height-weight relation, age, SBP, DBP, region, smoking status, etc. Moreover, *DiaVis* help to identify the most significant and high impacting factors of diagnosing diabetes disease. I used a user analysis and expert interviews to show the efficacy of this method on a real-world assertion dataset. In summary, the following **key contributions** are made by the work as follows:

- I present an observational study focusing on the domain requirements for analyzing customer behavior in understanding business risk.

- Visual interactive system named *'ExVis', 'MHIVis'*, and *'DiaVis'* are designed and presented that allows stakeholders and researchers to understand customer behavior and justify policyholders claim benefits.
- By interweaving the bayesian network, pattern mining, and statistical analysis with interactive visualization, I explore and inspect behavioral sequences from the life insurance claim records.
- I examine tasks of business analysts who aim to understand diverse customer behavior and visualize the outcomes.
- I provide some critical insights into the underwriting of policyholders for insurance managers to help making decisions.
- I also provide some concrete design implications for stakeholders derived from the findings.
- I report a user study with a large dataset and measured professional statements, which show the strength and usefulness of *'ExVis', 'MHIVis', and 'DiaVis'*.

4.2 Preliminary on Exploring Customer Behavior

4.2.1 Explainable Visualization System for Risk Management

This section describes the study on risk management in the insurance industry, VIS for risk management, and the explanation method for the VIS, respectively.

4.2.1.1 Risk Management in Insurance Industry

Risk management (RM) is the act of identifying, evaluating, and prioritizing risks (the result of an undesirable event), mitigating them, and maximizing an organization's ability to realize its potential. A risk management strategy that is effective will assist you in avoiding potential dangers and mitigating their impact. For instance, Islam et al. [120], focused adverse selection on a policyholder, diminishing an insurance company's effectiveness and profits in markets. Significant challenges for an IM are maintaining track of policyholders' complex behavior patterns and their varying levels of information. They proposed a association rule learning-based technique (ARLAS) for identifying policyholder behavior by utilizing an unsupervised learning method and analyzing 31,800 policyholders' life insurance data. Coussement and De Bock [67] studied on churn prediction of online gamblers to improve customer risk management (CRM) strategies. They gather customer data with the goal of optimizing customer churn prediction. They do so by employing ensemble learning techniques rather than single algorithms in order to assign high churn probabilities to actual churners. Additionally, Muller et al. [189] applied a Bayesian networkbased recommendation system for determining the clinical decision support system to assist risk management in making crucial decisions based on patient-specific evidence items such as medical tests and treatment history.

4.2.1.2 Visual Interactive System for Risk Management

To analyze data and identify its very subtle trends, patterns or context already recognized in text format, data visualization (DV) is a very useful tool in this regard. Replacing boring tables, charts or graphs, DV enables users to represent data in a very compact and interactive way through maps, bars, pie, etc. Nowadays, DV is used to display more complex data, identify the risk, and make a decision exploiting very cognitive and communication power of human being in security, economy, healthcare and so on. For example, complicated clinical data set are observed and analysed by psychologists with data visualization approach for child developmental health [99]. Same way this approach is now extensively used in financial and business institutions as well. Additionally, Eppler and Aeschimann [85] proposed some robust guidelines and framework for careful operation. Huang et al. [112] proposed a a visual analytical framework to prevent fraud attack and to assist human cognitive process. Similarly, Leite et al. [160] investigated to prevent any harmful transactions through another visual analytic approach to integrate human analysis into this process. Their approach focuses to uncover fraud operations identifying unusual financial events and fine tune existing automatic fraud alert system. Moreover, networked-guarantee loan risk management examines groups of enterprises who support each other for financial security [197]. Therefore, to build a healthy communication between risk experts and decision makers, an interactive visualization could be an efficient solution to identify, understand and analyze the most relevant risks [85, 112, 197].

4.2.1.3 Explanation Method for Visual Interactive System

To gain confidence, a machine learning-based recommendation system (RS) should exhibit transparency, allowing decision makers to understand the recommendations' key influences and conflicting facts, together with their associated degree of certainty [190].

When addressing the explainability of tools, explainable recommendation seeks to develop interpretable models for greater transparency, and such models often immediately result in the explainability of outcomes when focusing only on the explainability of recommended results. I view the RS as a black box and design additional models to explain the black box's recommendations. There are many types of RS explanations in the current research, including: (i) text-based explanations, which offer text sentences to users and may be created using pre-defined templates or generated directly using natural language generation models. (ii) visual interpretations, in which users will get picture-based explanations of the entire image or a highlighted area of interest in the image. The majority of studies has preferred graphical representations over text to communicate to the user, since text stresses the user, uses more space, and certain suggestions may be neglected owing to their intricacy. Some other study use the post-hoc method to describe a RS such as a Bayesian network (BN). These post-hoc methods provide explanations for each RS's decision. As a RS, BN determines the significance of individual evidence items via sensitivity analysis, which involves calculating the cost of omitting each piece of information. A graphical representation demonstrates the efficacy of visualizing the significance propagation via each node of a BN [298]. It illustrates how the findings interact to form a decision about the inherent nodal relationships in BN through a graphical method.

4.2.2 Exploring Mental Illness of Customer with Visual Interactive System

I focus on the particular problem of understanding of policyholder's mental health illness, and how stakeholders solve problems using information visualizations. In this section, I discuss two parts of work that relate to this problem. First, I set the research gap by outlining existing research efforts to articulate the 'mental health illness' using a visualization tool in the life insurance industry. Second, I discuss two 'information visualization process' research efforts that I able to uncover work practices previously and briefly discuss how I drew in my work.

4.2.2.1 Mental Health Assessment

Globally, mental health illness (MHI) is responsible for 32% of years of disability [139]. It is the main reason for claiming injury compensation in Australia [222]. It can have a deleterious effect on the physical and psychological health and the productivity of the work of an individual [21]. There is no doubt that mental illness has brought huge challenges and controversy to the insurance industry. Assessing the mental health illness associated with any insurance applicant is a core business characteristic of the insurance market. In the case of life insurance, it has been shown that the quality of care for mental illness has not increased to the same extent as that for physical conditions [222]. One study in the United States reported that psychiatric language usually varied from doctor to doctor, which leads to confusion in the interpretation of the diagnosis of a specific mental illness [17]. These risk assessments, also referred to as underwriting, are key to ensuring the economic viability of the insurance industry [65]. Underwriting mental health conditions can be challenging due to their complexity, particularly when applying for disability benefits [17, 65, 128]. When considering the negative experiences of mental health customers in using for and claiming insurance benefits, it is necessary to consider the personal information and risk assessments of mentally ill patients in the community. Additionally, how these pieces of information make psychiatric patients inequalities for a long time in insurance application and claim outcomes [130]. However, sometimes stakeholders are confused about dealing with the complexities of mental health illness and observe which individuals are engaged, disengaged, and those in between. So, this work provides design implications of a dashboard for insurance administrators that visualize policyholder's data relating to support a better understanding of policyholder's mental health illness.

4.2.2.2 Visual Interfaces for Stakeholders

Previous work on visualizing MHI has often focused on supporting the stakeholders in exploring user's interest by visualizing thread structures [186]. Several tools were developed to provide a visual overview and allows stakeholders to navigate the mental health states of users [54, 60]. However, such visualizations do not analyze the intrinsic relationships between mental illness and age range with policyholder's professions. Moreover, the above works have a lack of interpretability for recommendations to make a decision. Therefore, how to visualize the policyholder's mental health behaviors and why certain information's are recommended is a timely concern.

4.2.3 Exploring Diabetes through Visual Interactive System

This section provides brief literature on diabetes disease by exploring several analyzing techniques. Additionally, the existing approaches have been reviewed in several aspects, with a complete comparison at the end of this section.

4.2.3.1 Exploration of Diabetes Disease

Many researchers worldwide have been working on using various techniques to predict, forecast and analyze diabetes disease in recent years. For example, to classify diabetes data, Christobel Y. A. et al. [58] applied a new class-wise K-Nearest Neighbor (CKNN) classification technique and found that it outperforms simple KNN in terms of accuracy, sensitivity, and specificity. Lee et al. [156] predicted fasting plasma glucose which has been used to diagnose diabetes disease. They compared Naive Bayes (NB) and logistic regression (LR) to determine which technique provides the best accuracy, wherever NB delivers the best accuracy than LR. Additionally, Dey et al. [75] suggested an architecture based on the higher prediction accuracy of a sophisticated machine learning algorithm to determine whether a patient has diabetes or not. According to their analysis, artificial neural networks (ANN) provide the most significant accuracy. Wang et al. [284] established an effective prediction tool, especially for type-2 diabetes mellitus (T2DM), and investigated the possibility of genetic risk scores (GRS) in rural adults using multiple classifiers. They combined with GRS, which provides incremental performance for T2DM.

A noteworthy research work has been conducted which discusses the summarization of several chronic diseases using machine learning techniques which include diabetes disease also [95]. Sarwar et al. [234] performed a predictive analysis and revealed which algorithm is best suited for predicting diabetes disease. They proved that SVM and KNN provide the highest accuracy for predicting diabetes. From the research mentioned above works, it has been identified that existing studies covered various advanced technologies to determine whether diabetes or not. Most researchers preferred DL or ML techniques, and very few have explored visual analytic techniques to assess diabetes disease.

4.2.3.2 State-of-the-art: Visual Analytics

In this section, I review various studies that inspired examples to develop new technologies and prominently flourished this sector [124, 193, 316]. However, several recent studies focused on visualization techniques for exploring the field of healthcare analytics. For example, Kwon et al. [149] proposed a visualization system named DPVis to explore disease progression patterns and to derive clinical insights. They conducted a design study with clinical scientists, statisticians, and visualization experts to look into chronic disease pathways, namely, Type 1 diabetes. Swaminathan et al. [259] explored the ontology visualization tools, assessed them to see if each method was appropriate for end-user applications. In their research work, ontologies are used for the case of diabetes diseases. Wong et al. [290] described the usability and feasibility of software application in clinical practice and utilized diabetes data as a baseline. They test the usability of diabetes data especially type 1 diabetes (T1D). Bhardwaj et al. [29] performed Hadoop-based analysis to visualize diabetes data using Tableau software. They explored diabetes case studies and proposed a comparison among Pig, Hive, and Tableau.

Mahan et al. [176] used VA to assess and comprehend the prevalence and impact of diabetes worldwide and in the United States. Their research identifies countries with a higher risk of diabetes, particularly in the United States, where early detection and prevention can save lives and reduce medical costs. Furthermore, Rohlig et al. [220] presented a user interface that incorporates numerous VA tools, considerably boosting the practical value of the VA approach for detecting diabetic neuropathy by unifying access to their free exploration of medical diagnosis. The most common long-term complication is diabetic neuropathy; they try to minimize the risks and emphasize the need to detect nerve fiber abnormalities as soon as feasible.

There are many challenges to providing an interactive visualization system, especially for the healthcare sector [117, 118]. More recently, Rind et al. [217] discussed that poor data quality and ambiguity are among the challenges in building visual analytics tools for temporal electronic health records. Furthermore, Zang et al. [316] designed and implemented an interactive visualization tool titled as *IDMVis* to support clinicians adjusting intensive diabetes management treatment plans and categorize the process when making diabetes treatment decisions.

4.2.4 Summary

In summary, the existing research on customer behavior, MHI, DM and VA in the insurance industry is limited, and there are very few studies that have designed advance dashboards to meet the practical requirement of the stakeholders. Additionally, a very limited amount of datasets have been used in the existing literature. To deal with aforementioned challenges, in this study, I have studied a different domain (life insurance policyholder's MHI) and created a new dashboard by analyzing ten years data of a total 31,800 policyholder's to visualize the MHI and to provide some recommendations to reduce company losses, increasing service quality, improving risk adjustments and allowing for the improvement of insurance premium policy.

4.3 Methodology

This section describes how *ExVis, MHIVis*, and *DiaVis* are employed as a designed dashboard to bridge the information gap by using more intelligence in the analysis process and dynamic volumes of information through visual representations and interaction techniques.

4.3.1 Methodology of *ExVis*

This section discusses how *ExVis* is used as a designed dashboard for examining visual explanation for risk management and the details of data gathering and processing as shown in Figure 4.4.

4.3.1.1 Data Description

I collected three different types of data from Australian insurance companies, including (i) questionnaire data- the questionnaire was extensive and detailed, containing $31,800$ policyholders and 834 questions ranging from personal, medical, family history, occupational details, lifestyle, etc. (ii) demographic data- the demographic data of each user has various attributes including Insurance ID, Gender, Age, Occupation, and Postcode; and iii) policyholders' claims - the claims dataset consists of $27,458$ claim records containing 37 attributes that occurred throughout the country. From the dataset, I identified several attributes as key factors for exploring customer behavior and the categories of these factors as shown in Table 4.1.

4.3.1.2 Study Requirements and Design

In the insurance domain, decision-making is generated based on insurance guidelines, but the knowledge and expertise of insurance managers play a crucial role. The guidelines consider policyholder-specific data, such as Insurance ID, Gender, Age, Occupation, and Postcode. Such data is available from various sources but primarily present in unstructured and unsorted forms, challenging decision-making. IMs analyze available information and integrate it with their experience to make the decision. This process automatically filters out all appropriate information entities likely to be impacted by the

Table 4.1: Characteristics of respondents

Risk Category	**Risk Factor**
Diseases	Ashma
	Diabetes
	Rel Diabetes
	Cancer
Accident	Travel
	Alcohol
	Smoking 12 months
	Smoking 5 Years

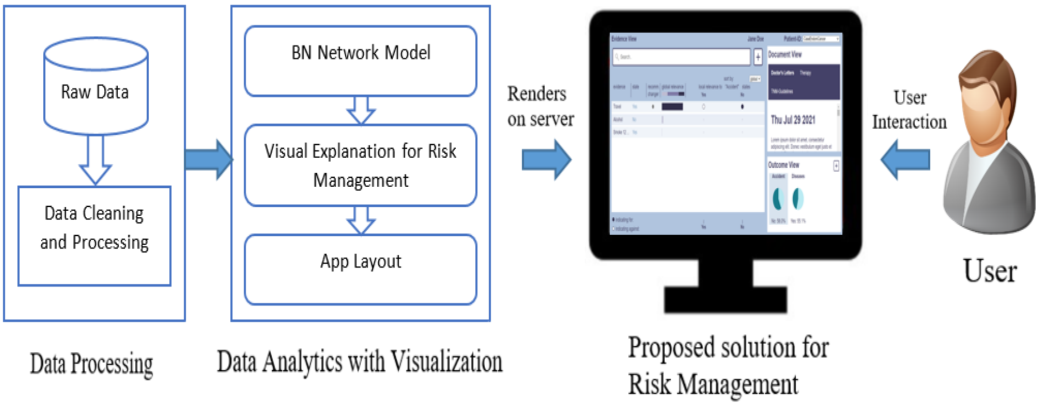

Raw Data
Data Cleaning and Processing
Data Processing
BN Network Model
Visual Explanation for Risk Management
App Layout
Data Analytics with Visualization
Renders on server
Proposed solution for Risk Management
User Interaction
User

Figure 4.4: The framework of *ExVis* system.

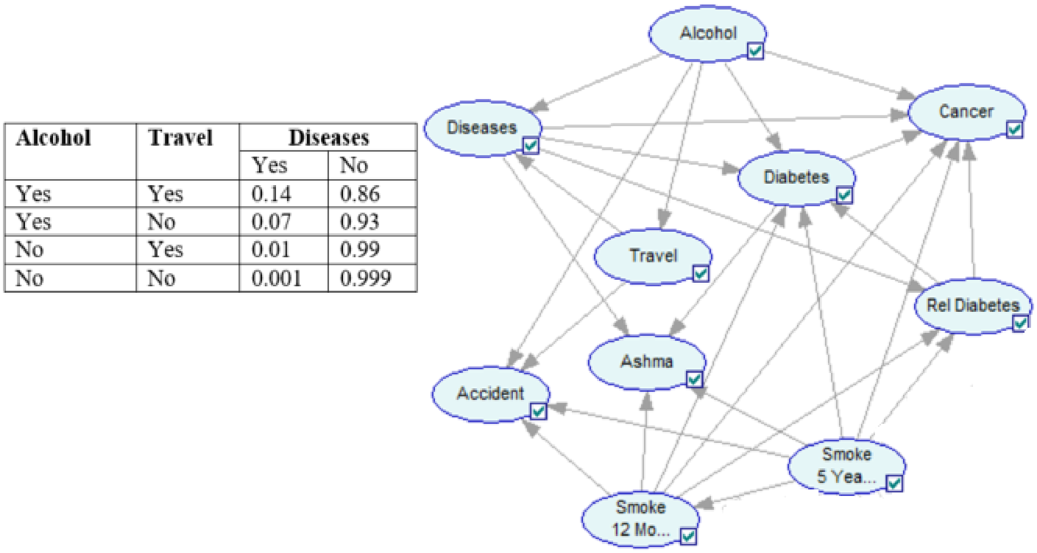
expected outcome. Therefore, considering the proposed design and functional criteria of *ExVis*, the approach must include representations of:

R1-Policy-specific data: IMs must be aware of facts that support and contradict the recommendation, as well as its potential alternatives.

R2-Policy Guidelines: IMs must be mindful of the diseases and accidental evidence used for generating a recommendation.

R3-Reasoning model: IMs must comprehend the underlying reasoning process for improving the acceptance and trustworthiness of a recommendation.

R4-Decision: it is necessary to demonstrate the accuracy of the computed recommendation. On the other hand, it must convey its uncertainty.



Alcohol	Travel	Diseases	
		Yes	No
Yes	Yes	0.14	0.86
Yes	No	0.07	0.93
No	Yes	0.01	0.99
No	No	0.001	0.999

Figure 4.5: The structure of bayesian network model.

Table 4.2: Conditional probability table (CPT) for bayesian network (BN).

Risk factors	Ashma	Dia.	Rel-Dia.	Cancer	Travel	Alcohol	Smo. 12 M	Smo. 5 Y
Ashma	N	N	N	N	N	N	N	N
Dia.	Y	N	N	Y	N	N	N	N
Rel-Dia.	N	Y	N	Y	N	N	N	N
Cancer	N	N	N	N	N	N	N	N
Travel	N	N	N	N	N	N	N	N
Alcohol	N	Y	N	Y	Y	N	N	N
Smo 12M	Y	N	N	Y	N	N	N	N
Smo 5Y	Y	Y	Y	Y	N	N	Y	N

4.3.1.3 Implementation Process

ExVis is implemented as a visual interactive explanation system as shown in Figure 4.4. I introduce my approach using a small Bayesian network (BN) for insurance risk management as shown in Figure 4.5. It consists of 10 nodes, such as alcohol, cancer, diabetes, rel-diabetes, travel, ashma, smoke 5 years, smoke 12 months, accident and diseases as shown in Table 4.1, and 22 relations. BN, a probabilistic graphical model, are being developed to assist people in making challenging decisions about uncertainty management. These systems are at the confluence of artificial intelligence, machine learning, and statistics. BN, developed based on Bayes' theorem, proposed by Thomas Bayes, is the basis for the Bayesian network. It calculates the posterior probability for a target of interest given a collection of input values usually referred to as the findings or evidence. It relates to inference on the target value. Based on the directed acyclic graph (DAG), which comprises a collection of nodes and directed arcs, statistical connections may represent directed arcs. Nodes indicate system variables, and an arc represents a cause-and-effect connection or dependencies between those variables in this graph, where node confidence values are calculated by combining conditional probabilities [229]. BN is resilient and compact as it can make sound predictions even when some factors are unavailable to its inherent nodal dependencies [298].

One of the critical concepts of BN is conditional independence which defines that given the values of its parents (PT), a variable X is conditionally independent of its non-descendants (ND) as follows:

$$P(X|PT(X), ND(X)) = P(X|PT(X)) \tag{4.1}$$

Another essential concept is the conditional probability table (CPT) which tabulates X's distribution for potential value assignments to its parents for each node X [298]. Thus, in this study, CPTs were developed by experts' group discussions as shown in Table 4.2. Their expertise and professions have been elaborated on earlier in this paper.

According to Muller et al. [190], global and local relevance is distinguished and determined for computing the cost of omitting each piece of evidence. First, Kullback-Leibler divergence (KLD) is used as a cost function to assess the dissimilarity between the target variable's probability distributions before (P) and after (Q) omission of evidence. The KL divergence is computed as an integral for continuous random variable distributions P and Q:

$$KLD(P||Q) = \int p(x) log(\frac{p(x)}{q(x)}) dx \tag{4.2}$$

Since I deal with discrete variables rather than continuous ones in the proposed system, the KLD is computed as the sum of P and Q, where P and Q denote the probability distributions of discrete random variables.

$$KLD(P||Q) = \sum p(x) log(\frac{p(x)}{q(x)}) dx \tag{4.3}$$

However, the KL divergence is not symmetrical because it doesn't follow the commutative law of two variables. Thus, the Jensen-Shannon divergence (JSD) measures the difference (or resemblance) between two probability distributions, which resolves the prior problem. It calculates a symmetrical normalized score using the KLD. This implies that P's divergence from Q is the same as Q's divergence from P:

$$JS(P||Q) = JS(Q||P) \tag{4.4}$$

I used Jensen-Shannon divergence (JSD) to quantify dissimilarity instead of cross-entropy based on the presence and absence of any evidence. The following formula can be used to compute the JS divergence:

$$JSD(P||Q) = \sqrt{\frac{1}{2}D(P||M) + \frac{1}{2}D(Q||M)} \tag{4.5}$$

where M is calculated as:

$$\text{and } M = \frac{1}{2}(P+Q) \tag{4.6}$$

JSD delivers a global level of relevance but none at the local level. As a result, I calculate the signed pre-post probability differences during the sensitivity analysis and then observe the shifts of its sensitivity.

4.3.2 Methodology of *MHIVis*

This section describes how *MHIVis* is employed as a designed dashboard for discovering, designing, implementing, and deploying as shown in Figure 4.6. Additionally, the details on data collection and processing are also included in the section for the representations of practical implications of the design.

4.3.2.1 Data Description

MHIVis work uses three different types of the dataset such as (i) policyholder's questionnaire data, (ii) policyholder's claim behavior, and (iii) SEIFA Index based demographic data are summarize in Table 4.3. The datasets have been collected from different data sources. First, I collect the questionnaire-based data used in this paper from one of the local life insurance companies in Australia. The questions and user's information were

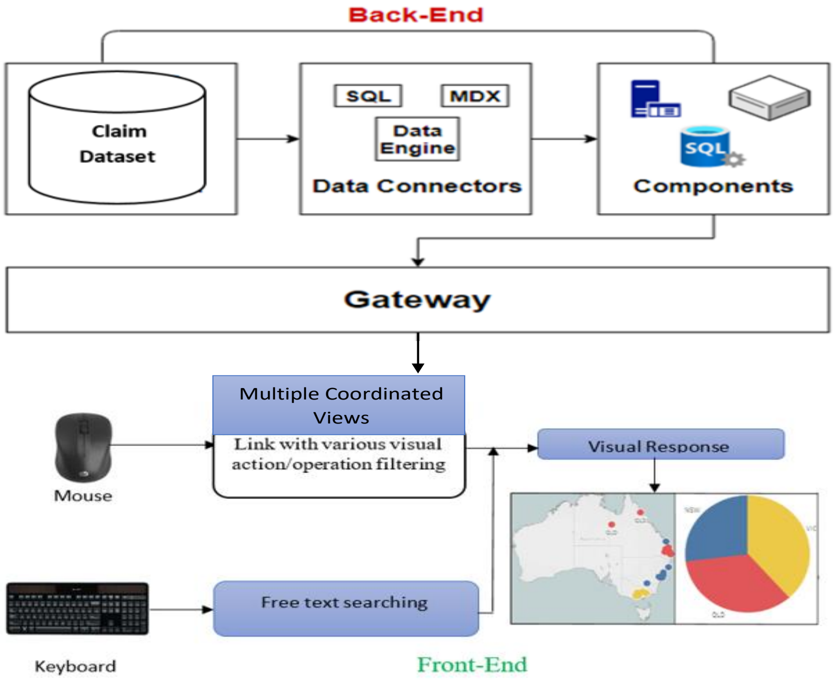


Figure 4.6: The framework of *MHIVis* system.

arranged and stored into the N*M matrix where rows indicate the user's information and columns represent the different questions. The primary dataset contains a total of 31,800 rows and 8,34 columns. Second, I collected policyholder's claim behavior data from the same insurance company. Third, to estimate the prior potentials, I collect the demographic information such as age, sex, profession, location, population, etc. of each policyholder's. Besides, as part of the demographic information, I use SEIFA index dataset [20].

4.3.2.2 Design and Function of *MHIVis*

This is designed to support stakeholders for exploring the MHI of policyholder's. According to Hoque et al. [108], The design study focused on four-stage of the design framework:

1) Discover: It covers the needs, problems, and requirements of stakeholders.

2) Design: After reaching a shared understanding of the *MHIVis*, I explored the design space of text analytics with multiple coordinated visualizations to support policyholder's.

Table 4.3: Dataset description.

Features description	**Values**
Total number of questionnaire dataset users	31,800
Total number of demographic dataset users	53,793
Total number of claim dataset users	4,200
Total number of mental illness policyholder	139
Number of depressive, anxiety & stress policyholder	79
Number of neurotic & personality disorder policyholder	60

3) Implement: I developed a new dashboard for stakeholders where stakeholders could effectively compare mental illness data between different age and professional policyholder groups.

4) Deploy: To analyze insurance data for exploring MHI of policyholder's, I deployed the design dashboard as a tool.

4.3.2.3 Implementation Process

Free text search: There has been a recent surge of attention in data visualization research by free text search [57]. It responds to user queries by either creating a new visualization within an existing visualization in the dashboard. In this study, I created a 'free text search' parameter and a 'free text filter' calculated field that referenced that parameter was like this, CONTAINS(LOWER([FieldName]), LOWER([Free Text Search])). In this strategy, the system first search for the field name in the dashboard and provide the visual result. For example, given the query such as 'top state in Australia', 'number 1 state', 'top 1 state', 'number 2 state', 'top 2 state', etc., the system will find a list of queries from the data table and generate a result in the dashboard. Here, number/top 1 state meaning that the system will provide the MHI policyholders information of that state in Australia.

Multiple coordinated views: In this strategy, I particularly focus on how my system allows selection within multiple views of the dashboard and use the frame-based query actions to generate the answer. For example, if you want to find how many 59 years mental health policyholder's are in QLD and what is there profession?. So, the system will first find how many mental illness policyholder's are in QLD, then find how many there are at 59 years and finally will find out what is their profession.

Table 4.4: Diabetes dataset description.

Attribute	Type	Description
Age	Number	Age in years
Residence	String	Rural or Urban
Region	String	Different areas
Working status	Boolean (yes/no)	Whether any of them is unemployed
Smoking status	Boolean (yes/no)	whether any of them is non smoker
Systolic blood pressure (SBP)	Number	SBP in numbers
Diastolic blood pressure (DBP)	Number	DBP in numbers
Height	Decimal	Height in decimal number
Weight	Decimal	Weight in decimal numbers
Class	String (diabetic or control)	Whether the objective identified as diabetic or normal

4.3.3 Methodology of *DiaVis*

This section explains how *DiaVis* can be used as a custom dashboard for finding, designing, showcasing, and deploying applications. In addition, information on data collection and processing is included in the section for representations of practical consequences of the design as shown in Figure 4.7.

4.3.3.1 Research Design

Qualitative research is the process of discovering the truth, finding a solution, and gaining knowledge. Qualitative research is usually made up of conceptual ideas and convictions that are used to investigate a specific topic [68, 310]. In comparison to policy reports and books, the case study method is brief, descriptive, and popular. For a deep, extensive, and intense analysis, the case study method is applied. Furthermore, the case study approach is ideal when the researcher has perfect cases to investigate and provide a thorough comprehension of the phenomenon [303]. Additionally, the prevalence of DM and its associated factors were used as a unit of analysis to determine the socio-demographic determinants that influence diabetes mellitus prevalence in Bangladesh.

4.3.3.2 Data Collection and Processing

In this study, I used the publicly available data where there were 1564 individuals people with nominal and ordinal variables [180]. From the dataset, I observed that two patients

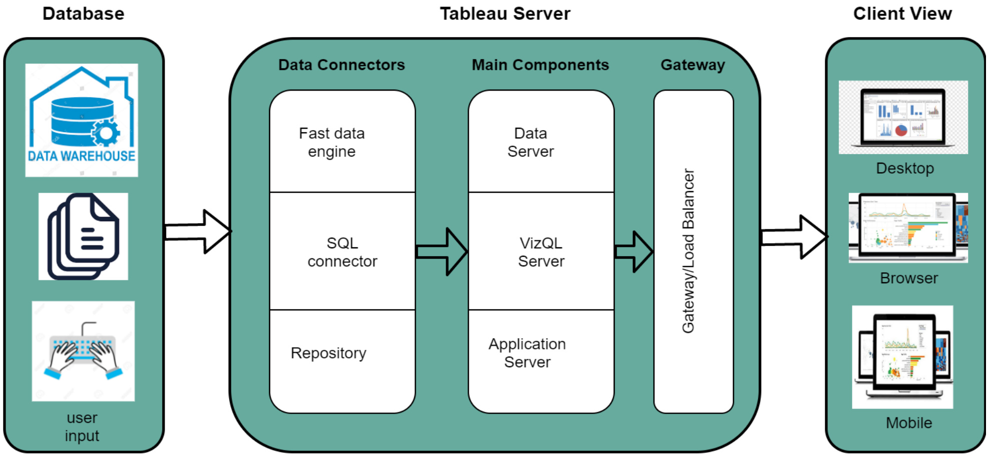


Figure 4.7: The methodological framework of *DiaVis*.

have zero cm of arm circumference, three patients have zero systolic blood pressure, two patients have zero diastolic blood pressure, and two patients have zero kg of weight in this dataset. It is noted that missing or null values are used to indicate zero values. Thus, in the data preprocessing phase, i must remove the null values. As a result, i have 1555 individuals' data in the final dataset, where there are 132 diabetes patients and 1423 healthy controls in the sample. Table 4.4 shows the attribute descriptions as well as a quick statistical overview.

4.3.3.3 Conceptual Model for the Study

Table 4.5 showing the percentage of the Sex: 73.74% male, 0.26% female, Age: 33.24% (35-39), 32.35% (40-44), 34.39% (45-49), Region: 21.93% Dhaka, 13.30% Chittagong, 12.15% Khulna, 9.53% Barishal, 16.30% Rajshahi, 14.96% Rangpur, 11.83% Sylhet, Working Conditions: 99.04% Yes, 0.96% No, Smoking Conditions: 15.86% Yes, 84.14% No, SBP: 48.39% (72-112), 50.57% (113-163), 1.02% (164-210), DBP: 81.96% (36-86), 18.03% (87-129), Weight: 35.26% (35.4-50.9), 48.71% (51-65.9), 13.97% (66-80.9), 2.02% (81-97.5), Diabetes: 8.57%, and Control: 91.43%. Figure 4.8 illustrates the conceptual model of this study. It shows various conditions for diabetics that are most potent and impact people's health and affect people's finances.

4.3.3.4 Requirements and System Design

I need to understand the typical decision-making process inside a clinical routine to create an effective visualization system for investigating diabetes. Also, I need to know what kinds of information concerning the decision problem are available and how humans

Table 4.5: Characteristics of respondents.

Characteristics	**Count of Responses**	**Percentage (%)**
Sex		
Male	1,560	99.74
Female	4	0.26
Age (years)		
Median age (range) = 43 (35 - 49)		
35-39	520	33.24
40-44	506	32.35
45-49	538	34.39
Region		
Dhaka	343	21.93
Chittagong	208	13.30
Khulna	190	12.15
Barishal	149	9.53
Rajshahi	255	16.30
Rangpur	234	14.96
Sylhet	185	11.83
Working Conditions		
Yes	1,549	99.04
No	15	0.96
Smoking Conditions		
Yes	248	15.86
No	1,316	84.14
Blood pressure SBP (Minimum = 72, Maximum = 210)		
72-112	729	48.39
113-163	144	50.57
164-210	16	1.02
Blood pressure DBP (Minimum = 36, Maximum = 129)		
36-86	1,282	81.96
87-129	282	18.03
Weight (KG) (Minimum = 35.4, Maximum = 96.5)		
35.4-50.9	540	35.26
51-65.9	746	48.71
66-80.9	214	13.97
81-97.5	31	2.02
Diabetes	134	8.57
Control	1430	91.43

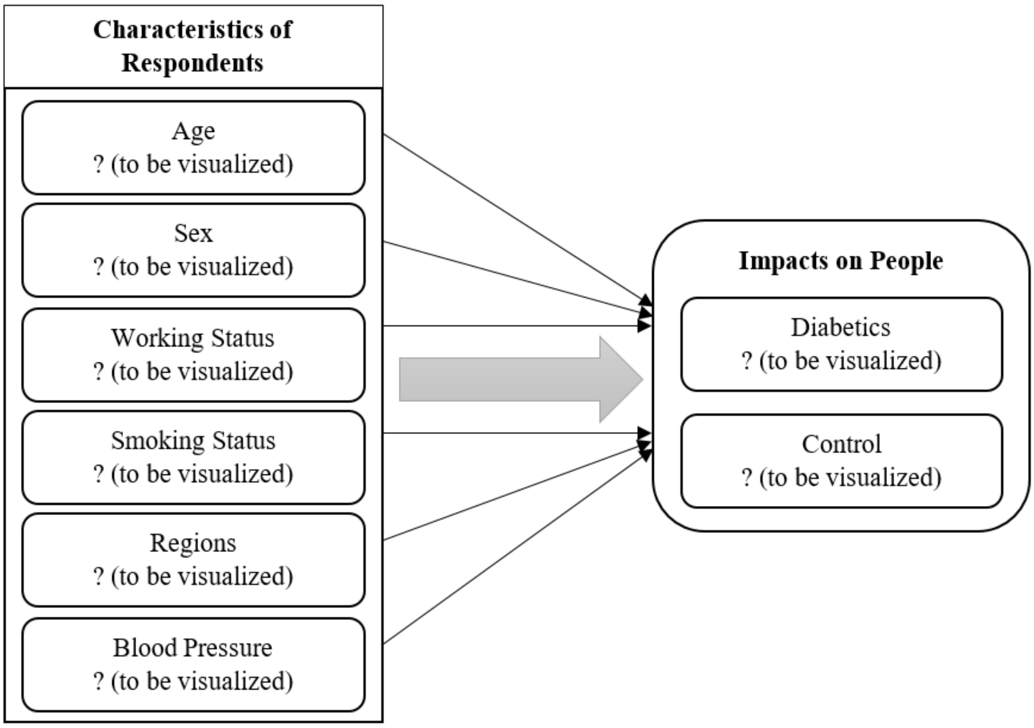


Figure 4.8: The conceptual illustration of the study.

process these data elements. Clinical decisions are based on clinical guidelines as well as the physicians' knowledge and experience. Height, weight, SBP, DBP, gender, smoking status, job status, and age are all considered in the guidelines. These data come from various places, and they're often unstructured and unsorted in terms of their relevance to the decision. Thus, based on expertise and in light of newly stated design and functional requirements for *DiaVis*, my approach includes the following representations:

R1: The visualization system makes it easier to assist decision-making regarding diabetes of a patient.

R2: Multiple coordinated views present more effortless and understandable data to identify diabetes patients and control patients based on different criteria.

R3: The mapping and region view is added to describe the number of diabetes patients in a region.

R4: Users must know the association between systolic blood pressure and diastolic blood pressure.

R5: Users require to understand the relationship between the height and weight of a person.

Figure 4.7 illustrates the methodological framework of *DiaVis*. It is designed to guide healthcare administrators (HA) in exploring the diabetes conditions of people. According to Islam et al. [121], the following four steps of the design architecture are: **1) Discover:** first, I discover the problems, challenges, and requirements. **2) Design:** After finalizing the understanding of exploring diabetes conditions of people, I explored the design space of qualitative analysis with multiple coordinated visualizations. **3) Implement:** I developed a new dashboard named "*DiaVis*", where HA could effectively compare diabetes condition data between different age, region, blood pressure, and occupational people groups. **4) Deploy:** by analyzing several case studies and corresponding refinements, I deployed the *DiaVis* system as a tool.

4.4 Experimental Analysis

4.4.1 Experimental Analysis of *ExVis*

ExVis computes a recommendation for the most rightful claim, "Accident and diseases." The findings of the system for visualizing the RM information of policyholders are discussed as follows:

Evidence view: The evidence view shows all evidence items organized according to their level of relevance for the computed outcome, e.g., a “Travel, Alcohol, Smoke 12 Months" recommendation.

Document view: The document view provides accident and disease-relevant information through the policy guidelines.

Outcome view: The computed probability distributions and a configurable set of accidental and disease scores are displayed in the outcome view to show the recommendations.

Network view: The network view is exposed by selecting evidence from the evidence view. This view is essential when IMs analyze the influence of new evidence and when they disagree with the evidence’s outcome.

4.4.2 Experimental Analysis of *MHIVis*

In this section, competing tools, and the findings of the dashboard for visualizing the MHI information of policyholder’s are discussed respectively.

Table 4.6: Feature comparison.

Features of dashboard	VMI	NHM	NHA	HIV Atlas	GBD	**MHIVis**
Visualization with global map	✓	✓	✓	✓	✓	✓
Visualization within age group	✓	x	x	x	✓	✓
Visualization with occupation category	x	x	x	x	x	✓
Review previous status across time/year	✓	✓	✓	✓	✓	✓
Free text searching with multiple views	x	x	x	x	x	✓

Notes: (✓) indicates the presence of feature selections and (x) indicates that it does not visualize the feature outcome.

Tableau allows the stakeholders to investigate the relationships within a plethora of graphs in the dashboard. During the initial exploration of the dataset, I aimed to understand whether mental illness is influenced by the categories in the variables. To understand the potential utility of the system, in Table 4.6, I compared MHIVis with five different visualization tools, including VMI, NHM, NHA, HIV Atlas, and GBD. These five tools are chosen because they also visualized information in the relevant areas, i.e., VMI in discovering mental health problems which include depression, drug, anxiety, and alcohol, etc. around the world, NHM in visualizing Nigeria health conditions with a limited dataset, NHA in visualizing national health account around the world, HIV Atlas in exploring the burden of HIV with a visualization scenario, and GBD in comparing different situations of MHI. Although existing tools are designed to visualize various information in the dashboard, to have a fair comparison of existing tools, I designed a new dashboard, including free-text searching with multiple coordinated views.

To identify and investigate the group of mentally ill policyholder's, IMs need to have a comprehensive understanding of policyholder's personal information. I design a dashboard that is user-friendly, simple, also in-depth to allow a thorough exploration of the MHI of policyholder's in Australia. This dashboard could be easily used by IMs. It will also facilitate to search a text query in a box and shows the result with multiple coordinated views within a dashboard. Furthermore, it gives both a zoomed-out and zoomed-in services of the Australian mental health landscape, such as the details of the mental disorders in different postcodes. As shown in Figure 4.2, explores the MHI of Australia as a whole, with details regarding Australia's spending. The map itself can be further zoomed in to see different postcodes, regions, and states in more detail.

The analysis in Figure 4.2 is necessary for IMs in designing appropriate policy packages for the future. For example, Indoor sedentary, Qualified professional, and Heavy Trades policyholder's are tending to claim for more benefits. Additionally, I outlined the postcode of all MHI policyholder's per state and show the geographical visualization for the different states in Australia. I can see VIC, NSW and QLD have the highest patient count, while Tasmania has the lowest patient count. However, the geographical size of MHI information may provides a misleading argument because the indigenous population has a much higher chance of suffering from MHI. The analysis of 'state information with MHI policyholder's by age group' and 'state information with MHI policyholder's by profession' suggests that IMs should pay more attention to the (25-54) range of policyholder's in NSW and VIC. In particular, NSW and VIC are the most popular location because they are highly developed states in Australia. Policyholder's tend to stay in these states, which can generate considerable benefit them. QLD is also a high-potential destination, where high-risk policyholder's tend to spend a long time. Therefore, IMs can investigate and develop appropriate business strategies among policyholder's when they provide insurance service facilities and reduce insurance loss.

4.4.3 Experimental Analysis of *DiaVis*

In this section, I provide the design and implementation of *DiaVis* - an interactive visual system to assist the users in understanding, engaging, and investigating the system thoroughly. Figure 4.3 shows how the system serves the user to interpret the diabetes dataset using a dashboard. The system's interface contains the search box, the regional map, and the other attributes of the dataset. Users can provide input to the dashboard using a keyboard or a mouse. From this figure, I can observe the number of diabetes cases in Bangladesh in the dashboard. The final dataset has 1564 patient information, where only 134 people are diabetes patients, and the number of control patients is 1430.

There are several inputs in the system. To visualize the number of diabetes cases, users can select their input choice according to the significant factors, including residence, region, age, smoking status, and working status. Moreover, the interface of the system produces four interrelated views: the mapping view, the region view, the Correlation Between SBP and DBP, and the Correlation Between Height and Weight are shown in Figure 4.9, 4.10, 4.11, and 4.12 respectively. I can observe the difference in the results through these views by changing the above-stated inputs. The details of these views are discussed in the following sub-sections:

4.4.3.1 Free Text Search

Free text search has recently sparked a surge of attention in data visualization studies. [57]. It responds to user requests by either developing a new visualization or incorporating an existing visualization into an existing visualization dashboard. I built a ’free text search’ parameter and a ’free text filter’ calculated field that referenced that parameter in this research, which looked like this: CONTAINS (LOWER([field name]),’ LOWER([field name]).’ The system under this method searches the field name in the dashboard first and then displays the visual result. For example, given the query, select single region type: ‘Dhaka or ‘Rajshahi, to select particular residence type: ‘urban’ or ‘rural’, to choose particular case type: ‘diabetes’ or ‘control,’ etc. According to the search query, the system will provide the result from the data table and initiate an impact in the dashboard.

4.4.3.2 Multiple Coordinated Views

Multiple Coordinated Views (MCVs) is an experimental visualization strategy that allows users to look at their data in a variety of ways [218]. The goal is to identify information in many possibly different datasets with a diversity of components and types and make sense of them. I focus on how the system permits selection within various dashboard displays and uses frame-based query actions to generate the answer in this manner. For example, do you want to find how many participants have diabetes according to their age and region?. So, the system will first see how many diabetes patients are there at a specific age and also find out their part.

4.4.3.3 Mapping View

Figure 4.9 illustrates the mapping view of the visualization system. It represents a map of Bangladesh, where the regions are highlighted in different colors based on their respective numbers of diabetic patients. Using the drop-down menu for areas, residence, working status, and smoking status, I can visualize the number of patients on the map. This map is also click-activated; by clicking on a region, the total percentage of diabetes patients for that region can be observed.

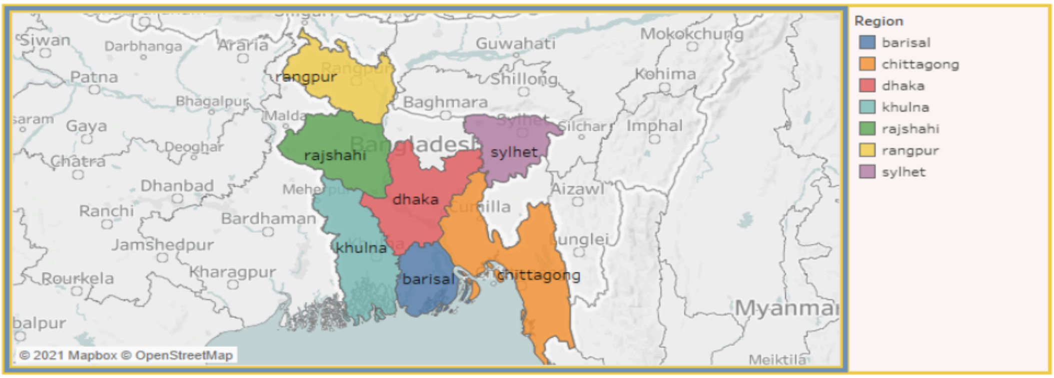


Figure 4.9: The mapping view of respondents.

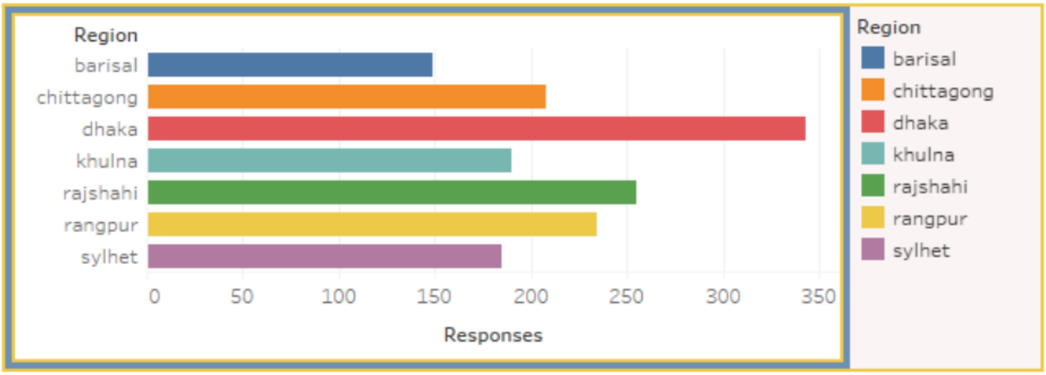


Figure 4.10: The regional view of respondents.

4.4.3.4 Region View

Figure 4.10 shows the regional view of the system where the number of responses for each region is presented using a bar chart. This figure highlights a hover-activated bar chart map, where the bar plots change based on the hover data for respective regions, which also changes the mapping view. I can also use the other drop-down and range slider input, which features the attribute "age." The fundamental of this view is a side-by-side comparison of the number of diabetes patients in every region. This aspect also provides perception as to how different regions have caused the production of diabetes patients.

4.4.3.5 Correlation Between SBP and DBP

In figure 4.11, I provide the association between SBP and DBP. I also present a scatter-plot diagram to visualize the sbp and dbp of diabetes and control person. The user interaction in this view is click-data enabled. The hover information, including sbp, dbp and diabetes, are displayed on the graph. The scatter plot changes based on the input

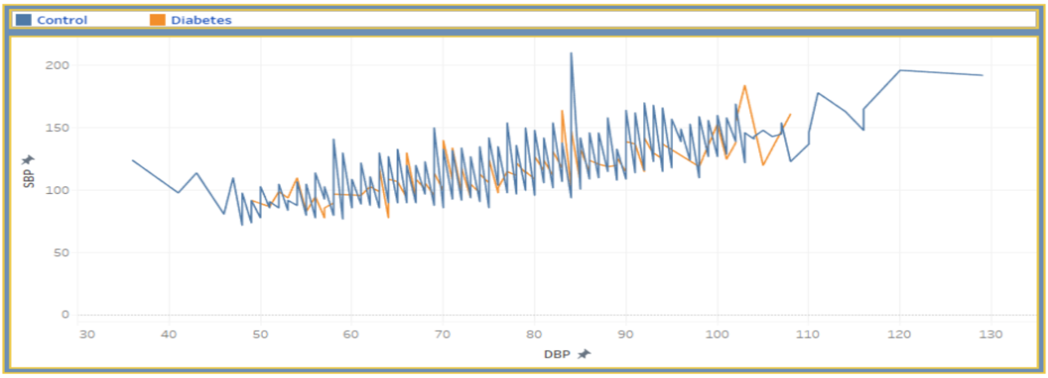


Figure 4.11: The correlation between SBP and DBP.

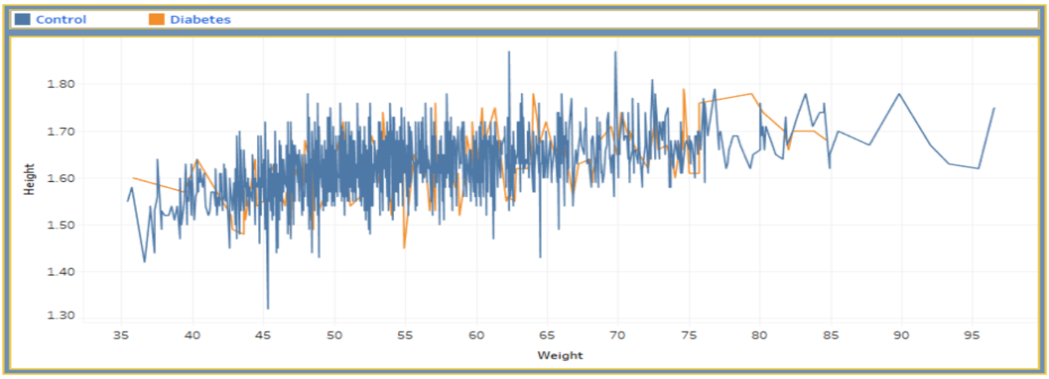


Figure 4.12: The correlation between Height and Weight.

data for respective regions, residence, smoking status, working status, which visualize the cumulative and single incidents.

4.4.3.6 Correlation Between Height and Weight

Similar to the previously explained view, Figure 4.12 presented the correlation between height and weight. This view also confers a scatter-plot diagram that reflects the association between a person's height and weight. This view is also click-activated. If I click on a point in the view, I can identify whether a person will have diabetes or not, based on the height and weight of that point. I can additionally notice the variations in the view by alternating the dropdown or slider inputs.

4.5 Result and Discussion

The model automatically learn to bridge the information gap by employing more intelligent in the analysis process, and dynamic volumes of information through visual representations and interaction techniques.

4.5.1 Result Analysis

To investigate and identify the group of mentally ill policyholder's, stakeholders, such as insurance managers (IMs), need to have a comprehensive understanding of policyholder's personal information. Therefore I design a dashboard to visualize detailed policyholder's of mental illness that perfectly contribute to discovering knowledge. The dashboard was designed to be user-friendly, simple, also in-depth to allow a thorough exploration of the MHI of policyholder's in Australia. This tool could be easily used by stakeholders, such as insurance managers (IMs). The design dashboard will facilitate to search a text query in a box and shows the result with multiple coordinated views within a dashboard. Furthermore, it gives both a zoomed-out and zoomed-in services of the Australian mental health landscape, such as the details of the mental disorders in different postcodes. As shown in Figure 4.2, explores the MHI of Australia as a whole, with details regarding Australia's spending. The map itself can be further zoomed in to see different postcodes, regions, and states in more detail. Additionally, Figure 4.2 reveals a more detailed image of Australia through the use of exploring the mental health risk of regions, postcode, profession, and age group.

The analysis in Figure 4.2 is necessary for insurance managers in designing appropriate policy packages for the future. For example, Indoor sedentary, Heavy Trades, and Qualified professional policyholder's are tending to claim for more benefits. Additionally, in Figure 4.2, I outlined the postcode of all policyholder's per state. I also examine the geographical visualization of mental health patients counts for the different states in Australia. I can see VIC, NSW and QLD have the highest patient count, while Tasmania has the lowest patient count. However, the geographical size of MHI information provides a misleading argument. It is noted that the indigenous population has a much higher chance of suffering from mental health issues. NSW and QLD have the highest amount of Aboriginal and Torres Strait Islanders. While this patient count may seem high, in comparison to their population, this is a low number.

As I used publicly available data on the Indigenous population, and thus, meaningful conclusions cannot be made. The analysis of 'state information with mental illness policyholder's by age group' and 'state information with mental illness policyholder's by profession as shown in Figure 4.2 suggests that IMs should pay more attention to the

(25-54) range of policyholder's in NSW and VIC. In particular, NSW and VIC are the most popular location because they are highly developed states in Australia. Policyholder's tend to stay in these states, which can generate considerable benefit them. QLD is also a highpotential destination, where high-risk policyholder's tend to spend a long time. Thus, IMs can investigate and develop appropriate business strategies among policyholder's when they provide insurance service facilities and reduce insurance loss.

4.5.2 User Study

Detail study was conducted to assess how to visualize customer behavior to gain a better insights. In order to investigate the potential usability and effectiveness of *ExVis, MHIVis* and *DiaVis*, I conducted a user study with five domain experts' opinions which helps us to understand the penitential usefulness and utility of the system. All participants have been 25+ years and had executive experience in diverse fields. The participants are primarily students and academic staff, where there are 3 males and 2 females who have expertise in visualization. They were familiar with the importance of customer behaviour in the business industry. In order to investigate the potential usability and effectiveness of *ExVis, MHIVis* and *DiaVis*, I asked participants to respond to some open-ended questions about the system are given in Table 4.7. The users actively participated and provided their comprehensive judgment. A qualitative user study is preferred, integrated by a quick survey, as I believed details insights would be strongest through in combination with a fair interview. The users' complete judgement was that the system would effectively understand, investigate, and recommend. Although nobody offered any compensation and tried to use the system earlier, I ensured them to access unlimited afterwards. I explored to determine possible usability and accessibility issues, examine what views of the system would, and they would not use, decide what facilities required the most value, and describe differences and missing capabilities. Finally, I observed that *ExVis, MHIVis, and DiaVis* could help the analysts to gain attention for specific risks management.

4.5.3 Discussion

4.5.3.1 Discussion on *ExVis*

I presented a decision making workflow of insurance claim, where I provided interactive solutions assisting in understanding and justifying the process of accidental and disease claim recommendation computation within ExVis. In this regards, I described several

Table 4.7: User study results.

No.	**Category**	**Question**	**Mean (μ)**	**Std. Dev (σ)**	**Min.**	**Max.**
Q1	Easy to use	*ExVis, MHIVis, and DiaVis* were easy to learn and use.	3.8	0.75	3	5
Q2	Insight	*ExVis, MHIVis, and DiaVis* was useful to explore insights patterns.	3.6	0.80	3	5
Q3	Insight	*ExVis, MHIVis, and DiaVis* allowed me to discover insightful queries about the data.	3	0.89	2	4
Q4	Essence	*ExVis, MHIVis, and DiaVis* helped me to generate knowledge about the claim data.	3.6	1.02	2	5
Q5	Speed	*ExVis, MHIVis, and DiaVis* enabled me to find interesting insights from the data quickly.	3.2	1.16	2	5
Q6	Confi.	*ExVis, MHIVis, and DiaVis* helped me to grow confidence about the interesting data insights.	4	0.63	3	5
Q7	Insight	I found that *ExVis, MHIVis, and DiaVis* were more interactive for exploring visual explanation.	3	0.89	2	4
Q8	Essence	*ExVis, MHIVis, and DiaVis* were easy and enjoyable to use.	3.6	1.02	2	5
Q9	Speed	*ExVis, MHIVis, and DiaVis* were worthwhile to explore explainable visual insights.	3.2	1.16	2	5
Q10	Essence	*ExVis, MHIVis, and DiaVis* allowed me to extract accidental and diseases information.	3.6	1.02	2	5
Q11	Speed	I feel confident to use *ExVis, MHIVis, and DiaVis* for discovering information from alcohol, smoke, cancer and so on.	3.2	1.16	2	5

visual assistance solutions addressing the associated needs of IMs at multiple steps of this workflow. The solution allows for building trust in the computed recommendation and for its application within policy guidelines as an additional objective decision-making opinion. The study also found that stakeholders want a complete dashboard that is more streamlined to match the needs and restrictions of their roles and work environments. In essence, I provide the following key insights:

- *ExVis* facilitates IM to make a decision.
- IMs may connect the BN's rationale to their own model and reasoning.
- Evidence items represent the key influences for the recommendation.
- Identifying policyholders' accidental and disease information can potentially be a very useful source for supporting IMs' goals;
- Future dashboard should be capable of managing bias visual presentations in broader cases;
- There may be some ethical issues surrounding dashboard design that may be required to protect privacy and data confidentiality of policyholders

4.5.3.2 Discussion on *MHIVis*

In essence, I provide several recommendations, such as a study should be conducted into the mental health of Indigenous Australians to understand their mental health risk better, availability of required large scale datasets to provide a more accurate and in-depth understanding of the Australian mental health climate. Although, earlier research reveals several possibilities of self-tracking data in mental health, however, in this study, I also highlighted the opportunities where self-tracking of mental health user can complement each other in order to fulfil their various unmet needs. The study also revealed stakeholders desire for comprehensive dashboard which is better streamlines to meet the requirements and constraints of stakeholders role and their work environment. I have the following key recommendations:

- QLD has the second-highest mental health risk and the highest population out of all the states. Thus, QLD needs to spend significantly more to help provide targeted support towards high-risk regions and postcodes.

- The study suggests that (25-54) range of policyholder's are considerably higher than others. So, IMs should pay more attention to policyholder's in NSW, VIC, and QLD.
- I noticed that the average age of the mental illness policyholder's derived between 35 and 45 years old. Therefore, the risk of mental illness claim behaviors for this population can be reduced by considering other factors as well.
- For larger dataset, behavior preference could be used to improve the performance of adverse claim selection.

4.5.3.3 Discussion on *DiaVis*

Based on the experiments and evaluations, I derive several points that enhance the performance of diabetes detection using the visualization system. I asked 20 participants to rate the system based on eight questions. The majority of participants ascertained the system to be useful, easy, enjoyable to handle and manage. They also perceived exciting insights about different factors behind diabetes and extracted knowledge about the disease. I have used nine factors in the system to detect whether the patient has diabetes or not. Moreover, I generated a heatmap to discover the correlation between the factors that affect the patient's insulin level and cause diabetes. From Figure 4.13, I can see that there is a strong correlation between systolic blood pressure and diastolic blood pressure which is 0.75. Moreover, there is an average correlation of 0.47 between height and weight. I deduce the following **key insights:**

- Age is a factor that interferes with the development of health needs during the life cycle. Among various significant parameters, the relationship between age and diabetes mellitus prevalence yielded inconclusive findings. Baquedano et al. [24], mentioned that older people have high diabetes risk. According to their findings, age is a critical component in providing diabetes education for disease management. Another factor to consider is the percentage of Mexicans who get type 2 diabetes before reaching 40. Diabetes was also found to be more likely as people got older, according to various research [133][141][53][83].
- DM is usually a hereditary disease. Obesity is one of the causatives and confounding factors that lead to the growth of diabetes mellitus. As a result, blood pressure is a very common complication of diabetes, affecting 20-60% of diabetic individuals,

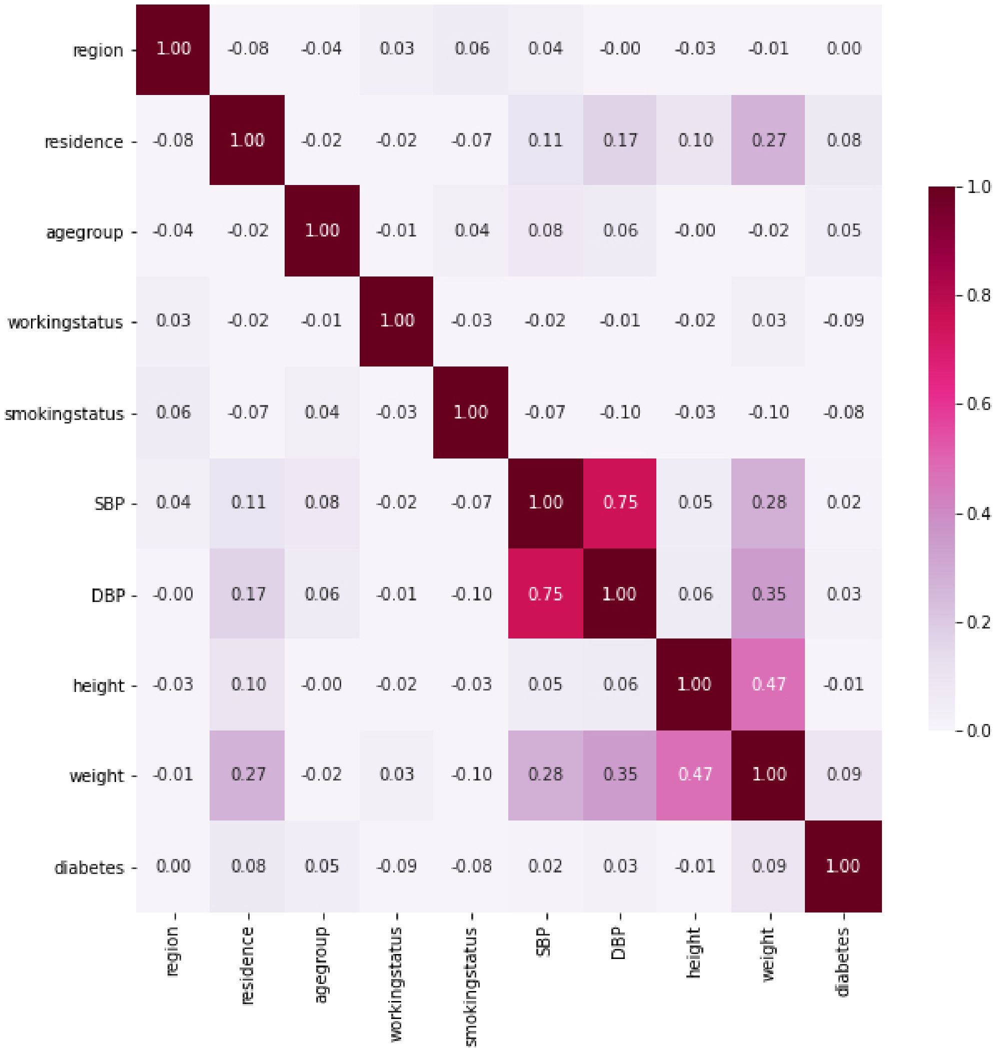


Figure 4.13: Similarity matrix of the attributes of *DiaVis*.

depending on weight, ethnicity, and age [12]. Additionally, diabetes-related comorbidities such as retinopathy and nephropathy are more frequent in hypertensive diabetic patients [15] [23].

- In Bangladesh, a study found that having a low educational level is significantly linked to poor glycemic control in people with Type 2 diabetes [3]. According to analysis, the level to which blood pressure should be reduced in a diabetic hypertensive patient, still unknown [225].

- The high prevalence of diabetes among middle-class Bangladeshi people in both urban and rural areas could be due to various reasons. South Asians and Bangladeshis have a significant prevalence of lifestyle-related chronic illness risk factors. For example, 43% of Bangladeshi participants in the INTERHEART study exhibited centripetal obesity [184]. In Dhaka, 58% had centripetal obesity, and 63% had a BMI of overweight or obese; the prevalence was exceptionally high among women (82% and 77%, respectively). In addition, 58% of men had smoked cigarettes in the past, and 34% were actively smoking [233].

- The correlations I discovered between diabetes, age, and body weight are identical worldwide. However, according to various research, those with hypertension have a higher risk of diabetes. Diabetes and high blood pressure complement one other since they have physical traits [12]. Diabetic patients have higher systemic blood pressure due to more excellent peripheral arterial resistance.

The results also show a relationship between multiple risk factors and considered three hypotheses. Another important finding of this study is the association between relevant cognitive functions in Bangladesh. Previous studies in developed countries have shown a strong relationship between poorly controlled diabetes and significant cognitive factors. However, no such research has been carried out in developing countries. Thus, the research study aids in the identification of the relationship between various factors and flourish how this relationship explores the analysis of diabetes diseases.

4.6 Summary

In summary, I present three interactive visualization systems (i) *ExVis*, (ii) *MHIVis*, and (iii) *DiaVis* to demonstrate insurance data for understanding behavior, diabetics, and mental health illness of customer. The study demonstrates (i) how the claim records can be understood by visualizing behavior sequences in a system. I described the strength and usability of *ExVis, MHIVis, and DiaVis* through a user study using a real-world claim dataset; (ii) how the system can help the stakeholders to utilize multiple interactions that complement together effectively; (iii) how the system helps decision-makers analyze policyholders' claims and allows for building trust in the computed recommendation and its application within policy guidelines as an additional objective decision-making opinion.

In essence, I provide the following **key insights:** (i) identifying mental illness policyholders can potentially be a beneficial source that not only helps to support the stakeholder's goals but also the governments; (ii) IMs may connect the BN's rationale to their mental model and reasoning; (iii) evidence items represent the critical influences for the recommendation.

In the future, the evaluation of various visual layouts for effective data communication could add because it is essential to consider when creating a visualization dashboard. It mediates individual differences' effects on performance and facilitates a viewer's understanding of visualizations. Additionally, several HCI evaluation methods could be applied and compared to find the best solutions with insurance companies.

CHAPTER 5

NLI-driven-DV to Explore Customer Claim Behavior and Manage Risk

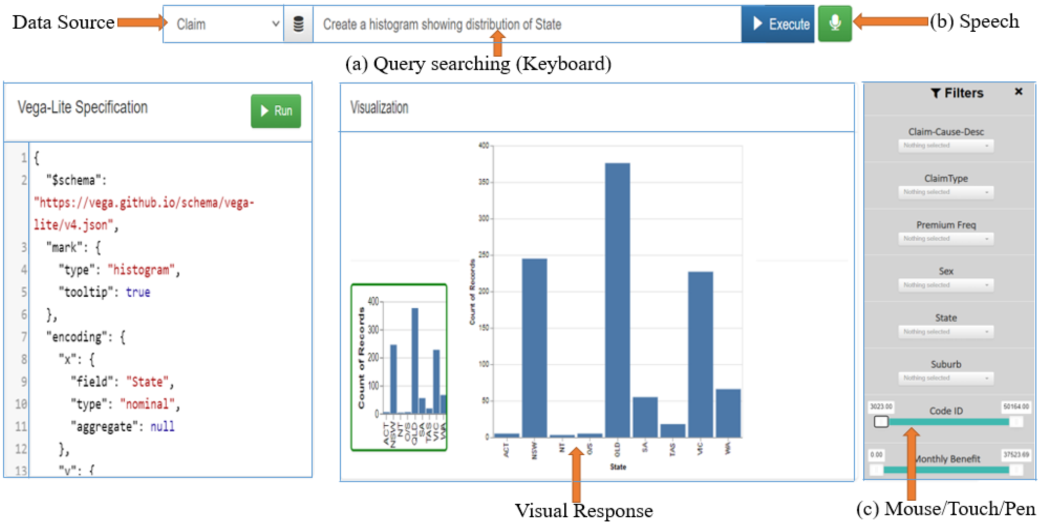
In this chapter, I provide the background and motivation in Section 5.1. Section 5.2 discusses preliminary work on customer behavior exploration, techniques, and behavioral data visualization followed by the detailed methodological discussion in Section 5.3. The description of the proposed visual analytic solution named (*InsCRMVis*) is presented in Section 5.4. I illustrate user study of the proposed solution to assess and discuss its capacity to inform the relevant variables for exploring customer behavior in Section 5.5. Finally, conclusions and future directions are provided in Section 5.6.

5.1 Background and Motivation

The financial industry is becoming more complex due to the lack of effective communication between risk experts and decision makers [112]. For example, a recent study of the life insurance industry in Australia found that managing risk involves more than protecting value [120, 121]. According to the Australian Prudential Regulation Authority (APRA), net claims expenses increased by 12.6%, i.e., from $22.1 billion to $24.9 billion from the year ended 2018 to the year ended 2019 . It is observed that re-insurers have a higher capital coverage ratio than direct insurers. Hence, insurance managers (IMs) need to take proper action to avoid fraud and reduce loss [85]. This strong control can be achieved through efficient communication between all the engaged bodies. Visualization is one approach to obtain such efficient risk-relevant information [143]. Although many forms of business diagrams such as tables, charts and formulas are a common solution for claim and risk management in the insurance industry, it may be challenging for IMs to identify the most relevant risks and to initiate adequate countermeasures. Therefore, data visualization is an effective solution to obtain risk-releted information for risk experts and decision makers [35, 121, 278].

Visual analytics solutions (VAS) are widely used for different purposes in a variety of areas such as finance, biomedical, education, forecasting research in academia etc. [1, 127, 160, 224, 264]. They enable researchers to gain better insights and to inform decision-makers through the analysis of large-scale datasets [45, 122, 269]. Moreover, a VAS expedites knowledge and provides evidence to improve outcomes [144]. For example, over the last few decades, several VAS have been proposed focus on fraud detection and customer monitoring [112, 159, 197]. These visualizations enable stakeholders to identify suspicious cases where traditional methods fail. However, most of the existing VAS are not effective in the insurance industry because (i) claim risk is very difficult to describe and extremely hard to visualize due to the multi-diversity of data; (ii) decision makers are not expert in the procedures of VAS with outcomes such as diagrams, risk maps, and the impact/likelihood positions of specific business risks. Moreover, existing research on the impact of customer behaviors on visualization processing has concentrated on primary insights that are employed in a non-interactive system (e.g., pie and bar charts), while the visualization outcomes on an interactive visualizations system are still limited. Thus, (i) to the best of my knowledge, there is no research on interactive visualizations system to systematically examine the visualization of insurance risk; (ii) interaction approaches only aid the theoretical processes for exploring risk visualization; and (iii) previous works only examine either low-level tasks (e.g., value retrieval) or high-level tasks (in specific analyses). Furthermore, there is no existing research which considers both high-level and low-level tasks together for decision making. Thus, with natural language interaction, the need to visualize and monitor the policyholders' claim risk is more urgent than ever before.

The recent development of natural language interfaces (NLIs) with data visualization has attracted immense attention from the research community, business decision-makers and industry to improve their net profit and considerably better performance has been achieved via predictive and analytic capabilities [250]. However, it is important to exam-



Create a histogram showing distribution of State

```
{
  "$schema":
"https://vega.github.io/schema/vega-
lite/v4.json",
  "mark": {
    "type": "histogram",
    "tooltip": true
  },
  "encoding": {
    "x": {
      "field": "State",
      "type": "nominal",
      "aggregate": null
    },
    "y": {
```

Figure 5.1: Summary of *InsCRMVis* interface components (Insurance claim and risk management with visual analytics). (a) query searching, (b) speech : allow users to freely express query to get visual insights, and (c) mouse/Touch/Pen: can be supported with visualization.

ine why NLIs are essential for data visualization and why they have been increasing in popularity? The existing studies on data visualization for NLIs found that NLIs have the ability to handle large data sets even with limited human and financial resources [84, 251]. It has been observed that existing NLIs have been used for the effective exploration and communication of ideas in various business domains [1, 90]. Additionally, NLIs can assist users to run queries to gain insights into large databases. However, when people want to query their data, they can have difficulty in generating the desired visual response using the existing NLIs. Moreover, existing NLIs are often intended for domain experts and have complex interfaces, hence challenges relating to ambiguity still remain. Therefore, an appropriate NLI with data visualization is required (e.g., language-based, speech-based, touch-based, speech+touch, language+speech+touch) to understand and explain large-scale datasets, particularly for insurance claim and risk management decision support.

The main contribution of this chapter is to design a new visual analytic solution (VAS) named *InsCRMVis* using NLIs to visualize policyholders' claims and risks in the life insurance industry. The development of VAS for risk visualisation has three goals: (i) to demonstrate the scope of risk visualisation, that is, where and when it may be beneficial and should be recognised as a valuable tool for risk managers; (ii) to present a checklist of the most important aspects to consider when visualising risks or risk-related data; and 3) to demonstrate how to visualise hazards for risk management, communication, and risk-related decision-making. Figure 5.1 illustrates how the system enables IMs to express their questions and intents more freely to gain insights from a large database of 26,817 policyholders' insurance claims to better manage risk [153]. To do this, speech and touch interaction can be supported by visualization, which enables IMs to follow up on the current status of policyholders to handle risk [9, 174, 249, 265]. Then, I collected 169 questions from the insurance stakeholders and found that VAS correctly answered 69% of all the questions. For evaluation, I performed a study of 10 users with three datasets, namely a questionnaire, demographics, and claims, using VAS. The experts' evaluation suggests that *InsCRMVis* can identify claim risks accurately assist IMs to reduce loss and guide changes to insurance premium policies for further development planning and management. Thus, the **key contributions** of this chapter are as follows:

- I introduce a new design space and present an end-to-end framework that enables experts to explore a large database of policyholders' claim behavior to reduce risk.
- I present the results of 26,817 policyholders' claim behaviors to expose the effects of different visualizations on understanding, distraction, driving performance, expert experience, and risk management.
- I collected 169 questions from the insurance stakeholders. I found that the system correctly answers 69% of these questions.
- To evaluate the performance, I performed a user study of 10 experts. The experts' evaluations suggest that *InsCRMVis* provides better insights and assists IMs to reduce loss and guide changes to insurance premium policies.

5.2 Preliminary on Exploring Customer Claim Behavior and Manage Risk

Natural language interaction for data visualization has been widely explored both by commercial software developers and the research community. I limit my discussion of the existing related studies to the following: (i) insurance claims and risk management; (ii) visualization for claims and risk analysis; and (iii) NLIs with data visualization. In the following section, I review the state-of- the-art in these areas to explain the motivation for the proposed VAS.

5.2.1 Insurance Claims and Risk Management

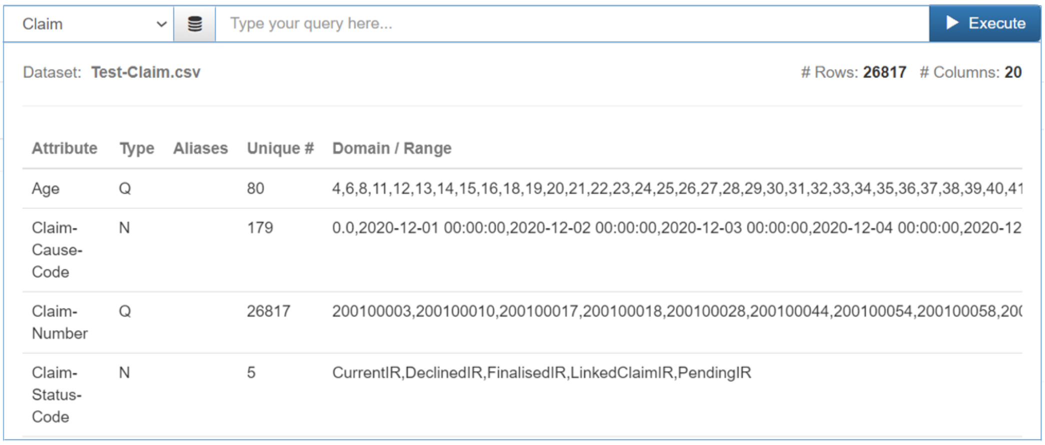
The exploration of insurance claims and risk management has attracted a significant amount of attention because a large number of policyholders have inflicted great loss on insurance companies and society as a whole [120, 276]. Insurance risk management is a branch of financial risk management and it includes life insurance and healthcare insurance [85, 143]. From the existing studies, it can be observed that life insurance claims and risk management has attracted more attention than other financial risk management issues. According to KPMG's life insurance insights 2020, Australian life insurance companies' premium revenue decreased by 6.1% to $17.3 billion, compared to approximately $18.4 billion per annum for the 2017 to 2019 period. Moreover, according to [18], approximately 21% - 36% of life insurance claims involve suspected fraud, but only 3% of perpetuators are prosecuted. Although researchers have expended great effort to address the problem of insurance claims and risk management using various effective risk management methods, these methods are often inadequate to handle claim and risk management problems [85, 120, 159, 197]. Moreover, existing studies have revealed that there is a need for data of better quality, consistency and transparency in relation to insurance claims [313]. Also, they lead to inferior outcomes in terms of extracting new insights to make a correct decision. Therefore, there is an increasing demand to improve risk management through the design and implementation of a cost-effective, practical, and real business-wide visual analytic solution.

5.2.2 Visualization for Claim and Risk Management

Claim and risk visualization employ systematic and interactive methods such as charts, maps, and conceptual diagrams to enhance the quality of risk communication along the entire claim and risk management life cycle. It helps experts and decision makers improve their understanding and deal more effectively with risk in the insurance industry. Visualization and visual analytics have been introduced both in academia and industry: (i) to provide a clear view of customers' adverse behavior, transaction monitoring, premium fluctuations, and in complex everyday decision-making [45, 99, 237]; (ii) to characterize data, user and task [42, 137, 254]; and (iii) to discover imbalances and monitor risk [71]. Whereas some contributions are domain-specific, e.g., visual animation is adopted to investigate the vast amounts of time-series data [16, 256, 279]. To monitor the behavior of a specific stock market user who has exhibited adverse trading patterns and to identify the real-time stock market performance, the 3D treemap is implemented [89, 112]. To detect adverse user behavior, the coordinated specific keywords visualization is developed within the wire transactions [247]. Additionally, various interactive visualization systems are developed to help the stakeholders make an immediate decision for different business scenarios [63, 309]. The clustering-based visualization system has been used in financial risk monitoring, discovering imbalances in financial networks, and for predicting head and neck cancer patients [4, 50, 268]. However, there are very few works on claim and risk management in the insurance domain. Moreover, the existing systems have limitations in relation to investigating a large number of variables and satisfying specific requirements, e.g., measuring new claims costs, number of accidental claims, and the number of mental health claims of domain experts. Therefore, in this study, I address the gaps in the existing visualization systems and meet the demands of domain experts.

5.2.3 Natural Language Interfaces with Data Visualization

Natural language interfaces (NLIs) are emerging as a promising paradigm for data analysis with visualization [117, 198]. It is gaining in popularity because it helps to improve the usability of visualization systems. Typically, these interfaces respond to user queries by either creating a new visualization and/or by highlighting answers within an existing VAS. It has been explored by the research community and also as commercial software. Existing studies have provided various NLI-based VAS that use well-structured commands to specify visualization. For example, NLI-based VAS such as articulate, and ConveRSE enable people to explore how NL affects in the incorporation of digital assistants and recommendation systems [115, 257]. DataTone manages ambiguity to let people specify a visual response through NL queries and to develop the useful NLIs for data visualization [90]. FlowSense allows the user to write a query and visualization components to specify system functionality [305]. Eviza incorporates a probabilistic grammar-based approach and a finite state machine to provide NLIs for an interactive query dialog [240]. Evizeon supports compound queries, and lexical cohesion with visualizations [109]. The ideas in Evizeon and Eviza were also utilized to describe the Ask Data feature to specify NL queries in an organized shape in Tableau. From the aforementioned systems, it has been observed that NLIs provide an opportunity to ask any questions in generating the desired visualizations using natural language. However, in the insurance domain, there are no NLI-based VAS to identify insurance claims and manage risk management. Therefore, inspiried by the aforementioned visualization systems, I leverage data visualization with natural language interactions to explore



Dataset: **Test-Claim.csv** # Rows: **26817** # Columns: **20**

Attribute	Type	Aliases	Unique #	Domain / Range
Age	Q		80	4,6,8,11,12,13,14,15,16,18,19,20,21,22,23,24,25,26,27,28,29,30,31,32,33,34,35,36,37,38,39,40,41
Claim-Cause-Code	N		179	0.0,2020-12-01 00:00:00,2020-12-02 00:00:00,2020-12-03 00:00:00,2020-12-04 00:00:00,2020-12
Claim-Number	Q		26817	200100003,200100010,200100017,200100018,200100028,200100044,200100054,200100058,20(
Claim-Status-Code	N		5	CurrentIR,DeclinedIR,FinalisedIR,LinkedClaimIR,PendingIR

Figure 5.2: The data sources of the policyholders claims.

insurance claims and manage risk.

In summary, the existing research on data visualization for exploring insurance claims and risk management is very limited. Although few studies have developed an interactive visualization system, there is no study on data visualization with NLIs to meet the practical requirements of risk domain experts in the insurance industry. Thus, to the best of my knowledge, this is the first work using NLIs with visual analytics approaches to address insurance claims and risk management issues.

5.3 Methodology

5.3.1 Data Description

This work uses three types of data collected from an Australian insurance company, namely (i) questionnaire data ; (ii) demographic data; and (iii) policyholders' claims. All the attributes of the questionnaire dataset are binary where the demographic and claim datasets consist of binary, categorical, numerical etc. data. The attribute descriptions are given in Figure 5.2. A brief description of each dataset is presented respectively.

Questionnaire dataset: I acquired the dataset, amassed over 10 years, from a screening questionnaire provided by an insurance company. The questionnaire was considerably large and detailed, comprising information on 64,000 policyholders from 834 questions ranging from personal, medical, family history, occupational details, lifestyle, etc. with responses labelled 0 for ‘No’ and 1 for ‘Yes’. For example, the questions asked whether *the participants drink alcohol or not*, whether *they have cancer or not*, whether *they smoke or not*, whether *they have a disease or not*, etc.

Demographic datasets: The demographic dataset comprises five attributes, namely insurance ID, gender, age, occupation, and postcode. The ‘gender’ attribute comprises ‘male’ and ‘female’. The ‘postcode’ attribute reports the Australian postcode of the policyholders’ place of residence. The ‘age’ attribute reports the age of the applicant in whole years, and shows the youngest applicant is 3 years old and the oldest is 78 years old. The ‘occupation’ attribute contains 18 different categories such as ‘T-Trades’, ‘S-Supervisor of Trades’, ‘R-Special Risk’, etc. As part of the demographic information analysis, I also use the Socio-Economic Indexes for Areas (SEIFA) data set.

Claim dataset: The confidential customer claim dataset is provided by the IMs for research purposes only. In total, 26,817 claims were recorded from 2010 to April 2019.

5.3.2 Data Pre-processing

As discussed in Subsection 5.3.1, various information is recorded in the dataset consisting of various attributes. Since attributes have values in different categories, the dataset may contain missing values. To simplify the system to ensure only the most significant data is used, data pre-processing involved reducing the less important and redundant attributes which offer no benefit to exploration and analysis. As part of the data preprocessing, redundant fields that were not eliminated were combined. Finally, the dataset comprised information on 26,817 policyholders with 21 attributes relating to insurance claims and risk management. I applied these cleansed datasets to provide a broader and more comprehensive analysis to explore claims and risk management in the insurance industry.

5.3.3 Domain Characterization and Design Consideration

To develop an effective visualization for exploring claim and risk management, I first must understand the common decision-making process within insurance guidelines. I need to know what types of information are available in the decision problem, and how humans process these information entities.

Table 5.1: Key questions identified in collaboration with domain experts.

RQ1	Why is an interactive visual analytic tool necessary for the insurance domain?
RQ2	What the key factors/content should be depicted when exploring insurance claims and visualizing risks/risk-related information?
RQ3	Who will monitor the visualization system to control risk?
RQ4	When stakeholders should consider an interactive risk visualization system a useful tool in light of the benefits it provides?
RQ5	How can natural language interfaces (NLIs) be supported through an interactive visualization system to investigate insurance claims and manage risk?

In this work, I collaborated with a team of IMs who have more than five years of working experience. The task is to understand the decision-making problem through a series of interviews and discussions. Therefore, I collected several questions that could not be answered by existing VAS as listed in Table 5.1. These questions suggest that analysis should be able to inspect the behavior of both individual and/or group policyholders, as well as identify the most important information for exploring insurance claims and risk management.

I note that the collaborators wanted to conduct a comprehensive analysis of policyholders' behavior and also wanted to find specific values and information. Thus, it was essential to occupy the insurance claim data without losing detail, e.g., being able to display specific values. As the capability to present response defined the demands for designing a VAS, the design efforts focused on bringing complementary views of various relationships and supporting IMs to examine representative variable in relation to adverse behavior. The analysis shows that the VAS has the capability to compare variables in terms of policyholders' behavior. In Section 5.3.4, I discuss the system's properties which are useful in obtaining responses to such queries.

Based on my experience and taking into account recently proposed design and functional criteria, the system *InsCRMVis* must include representations of:

R1 Domain-specific data: The key influences need to be emphasized and sorted regarding their relevance to the policyholders' claim benefits. Hence, the IMs must know the supporting as well as contradictory facts before making a decision on a claim for benefits and the potential alternatives.

R2 Key factors: IMs must be aware of the information provided by the policyholder in relation to a claim for benefits through visualization in risk management.

R3 Monitor and/or control risks: A conceptual NLI incorporating the purpose (why?), the content (what?), the target groups (for whom?), the situation (when?) and the format (how?) allows IMs to systematically explore data visualization in risk management and to discuss new insights.

R4 Decision: Insurance decision-making for claim and risk management aims at finding the right information, aims at finding an adverse outcome for a specific policyholders.

R5 NLIs for Investigating insurance claim and risk management: NLIs enhanced by visualization requires thorough task analysis and domain expertise to explore risk management and claim analysis

5.3.4 Visual Analytics Solution

According to the novel visualization toolkit named *NL4DV* developed by Narechania et al. [193], I design a proposed natural language based framework named *InsCRMVis* for risk data visualization. Figure 5.3 illustrate the components of the proposed methodology for VAS. I aim to cover the scope of risk visualization, that is to say, highlight various purposes, what are the contents and for whom risk visualization can provide benefits. I consider *InsCRMVis* to be a useful tool and provide a checklist of the key factors to consider when visualizing risks or risk-related information. *InsCRMVis* consists of four components: 1) data collection and processing; 2) designing a visual analytics framework; 3) applying the framework to a specific domain; and 4) evaluation.

Step 1: First, I collected a dataset for data processing, organizing, and cleaning, as described in Section 5.3.2. This ensures the dataset is effective, as organizing and cleansing data make it more reliable and free of duplication.

Step 2: Like many other web applications, the visual analytics framework named *InsCRMVis* consists of two components: 1) *InsCRMVis*- Automatic Query Answering, and 2) *InsCRMVis*- Multimodal System. The first component of the *InsCRMVis* framework allows the user to search various queries to gain insights into a large database and the

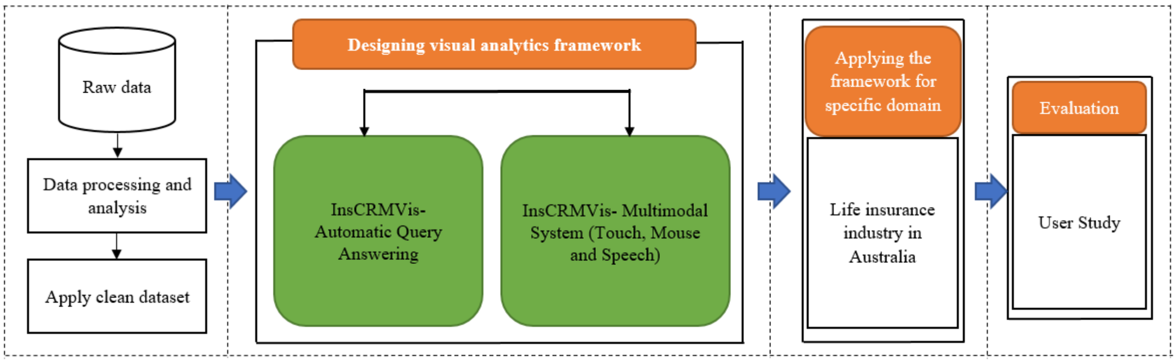

Raw data
Data processing and analysis
Apply clean dataset
Designing visual analytics framework
InsCRMVis- Automatic Query Answering
InsCRMVis- Multimodal System (Touch, Mouse and Speech)
Applying the framework for specific domain
Life insurance industry in Australia
Evaluation
User Study

Figure 5.3: Proposed architecture of NLI based visualization.

second component allows interaction between various plots in a visualization system through touch, mouse and speech. The panel also has a filter option based on the claim score.

Step 3: The *InsCRMVis* framework is applied to the life insurance domain in Australia and allows IMs to explore insurance claims and risk management.

Step 4: The domain-specific application framework is dependent on expert evaluation to obtain feedback to assist in reducing loss, and guiding changes to insurance premium policies.

As illustrated in Table 5.1, I provide specific questions relating to why, what, for whom, when and how the risk-related information should be visualized, as shown in Figure 5.4. Therefore, it is important to start with these questions which will provide possibly useful answers for risk visualization. Through the interface, I can observe this represents a process view of risk depiction; a solution that emphasizes the act of visualizing, rather than just the resulting graphic artifact.

5.4 Experimental Analysis

In this section, I present a new design space data visualization architecture namely *InsCRMVis* to explore insurance claims and risk management, as shown in Figure 5.5. The designed framework combines multiple visualization components such as text, speech, touch etc. which conveys the claim behavior of each policyholder in a consistent

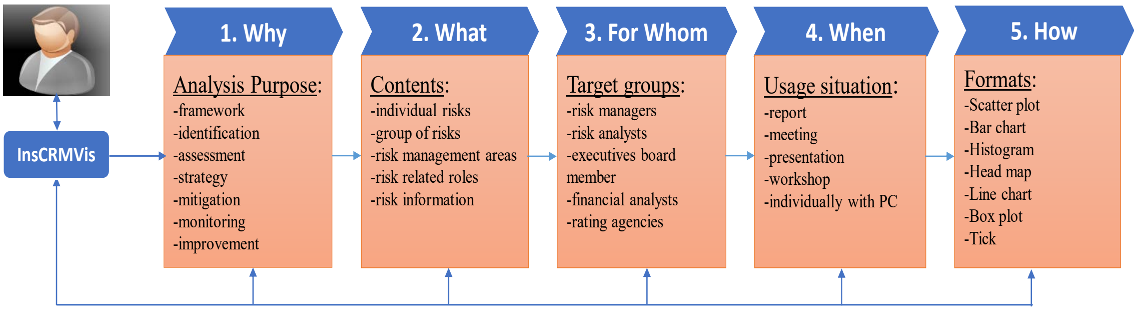


Figure 5.4: System overview: key questions of the risk visualization framework.

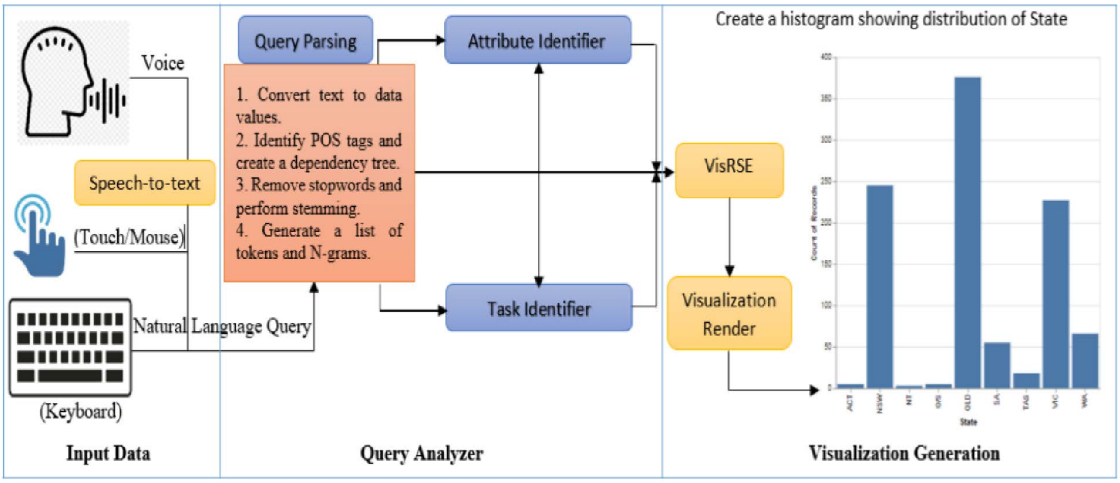


Figure 5.5: Framework: overview of the interface functionalities such as input data, query analyzer, and visualization generation.

representation of the data observations. The approach is similar to the method proposed by [193]. It integrates multiple natural language processing and visualization techniques into a framework to support risk experts in the investigation of the claim behaviors of policyholders. It comprises three key components, namely 1) data interpretation, 2) query analyzer, and 3) visualization generation. In the following, I briefly described how *InsCRMVis* uses these key components to to explore and minimize claim risk?

5.4.1 Data Interpretation

I use insurance claims along with questionnaires and demographic data to infer various types of attributes. For example, the dataset contains the attribute 'Monthly Benefit' with a range of values. When I look at temporal information, the system may provide misleading information which can lead to poor decision choices. Thus, to overcome this issue, *InsCRMVis* iterates through the underlying data item values to derive metadata consisting of the attribute types such as quantitative, nominal, ordinal, temporal along with values for each attribute in a range. This attribute metadata is utilized to interpret queries to analyze exact tasks and generate appropriate visual responses.

5.4.2 Query Analyzer

A natural language interaction-based visualizer should be able to analyze the phrases in the query that are more informative. To generate a visual response from a query, NLIs need to identify the related information such as analytic tasks, data attributes, type of visualization, and values as shown in Figure 5.6. For example, *'Create a histogram showing distribution of M Sex in NSW'*. In response to this query, *InsCRMVis* performs three operations: query parsing, attribute interface and task interface.

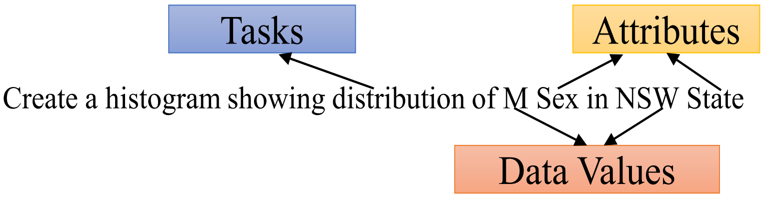

Tasks
Attributes
Create a histogram showing distribution of M Sex in NSW State
Data Values

Figure 5.6: An illustration of a query analyzer while interpreting NL queries.

In order to extract details and adopt more relevant phrases, the query parser first runs a set of NLP blocks that include part of speech (POS tags), dependency tree, and N-grams. Followed by query parsing, *InsCRMVis* searches for data attributes that are specified both explicitly and implicitly. Finally, *InsCRMVis* analyzes the remaining N-grams for references to analytic tasks such as correlation, distribution, derived value, trend, and a fifth filter task, as shown in Table 5.2.

5.4.3 Visualization Generation

InsCRMVis uses Vega-Lite to operate as the regulating visualization grammar to visualize up to three attributes at a time. It holds the Vega-Lite marks such as tick, bar, point, line, arc, area, boxplot, text and encodings: x, y, size, color, row, column, etc. [235]. Similar to *NL4DV*, the combination of Vega-Lite marks and encodings allows *InsCRMVis* to support a variety of popular visualization types like bar, histograms, line, strip plots, pie charts, box plots, area, and scatterplots [193]. To provide insights related to the query, *InsCRMVis* analyzes the query for explicit requests for visualization types (e.g., 'pie

Table 5.2: Types of queries and visualization observed in this study.

Query Example	Task	Visualization Type
Show a scatter plot of age and monthly benefit for policyholders under the age of 30.	Correlation	Scatter plot
Show me the relationship between age and claim cause description.	Correlation	Scatter plot
Show me bar chart of claim casue desc and ier-score.	Derived Value	Bar chart
Show me the distribution of irsad-score.	Distribution	Pie chart
Visualize monthly benefit for depression/anxiety of males in Australia.	Derived Value	Bar Chart
Show an average sum of amount for the state of QLD.	Distribution	Strip Plot
Create a histogram showing the distribution of State	Distribution	Histogram
Show a line chart of claims by state in Australia.	Trend	Line Chart
Show me premium frequency in the state of NSW.	Distribution	Histogram

charts', 'histogram', 'box plots') or implicitly infers visualizations from attributes and tasks. To implicitly determine visualizations, *InsCRMVis* utilizes a combination of the attributes and tasks derived from the query. Then, it compiles the inferred visualizations into a visList. Each object in visList is composed of a vlSpec containing the Vega-Lite specification for a chart, an inference Type field to highlight if a visualization was requested explicitly or implicitly derived by NL4DV, and a list of attributes and tasks to which a visualization maps.

5.4.4 Implementation

The *InsCRMVis* system is developed as a web-based application, where Python and Flask are used to develop the back-end to support data processing and analysis. JavaScript is used to implement the front-end where data-driven documents (D3) are used to build visualization views. A combination of HTML, CSS elements provide the interface and the AngularJS framework is used to structure the web application using a model-view-controller paradigm. The web-based front-end is connected to the back-end through a query engine interface where the query engine brings in aggregated data from the back-end based on interactions and user selections on the front-end. Figure 5.1 displays the primary screen of the *InsCRMVis* front-end which comprises a full view for visualizing insurance claim datasets.

5.5 Result and Discussion

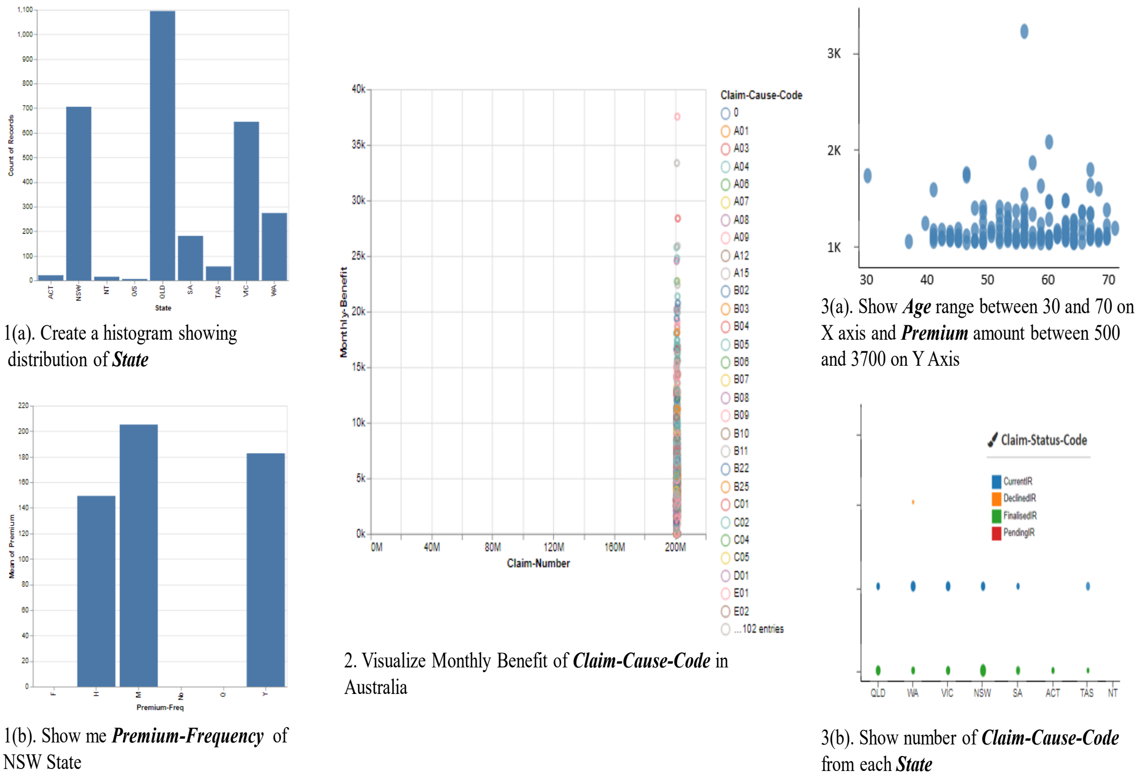
5.5.1 Result Analysis

InsCRMVis visually guides domain experts to identify claim behavior and reduce risks using the various functions described in Section 5.3.4. I highlight the following outcomes to the selected motivating examples raised by the questions listed in Table 5.1. In this study, I used 169 questions and I see that 69% of the answers generated by the system are correct.

Identifying and understanding relevant risks (Q1 and Q2): Figure 5.7 shows a variety of questions and charts generated by the system. I observe that the use of an appropriate VAS can help the stakeholders become aware of specific risks and provide way to deal with these risks adequately. For instance, an insurance company cannot reduce the number of risky policyholders directly. By analyzing different factors with a reliable system to establish a fair claims management process, a good overview of many relevant business decisions can be gained. Thus, the design of an interactive visualization system is important to avoid any complications with factors that are not relevant.

Exploring situations for risk visualization (Q3 and Q4): The exploration of a large database for risk visualization can provide useful insights for various risk-related purposes. In Figure 5.4, I provide target groups and usage situations to make sure for whom and when risk information needs to be visualized to make a decision. For example, IMs most likely want to identify risky users to allocate adequate resources to mitigation measures and to understand how their risks are interrelated. In Figure 5.7, I provide some question answering to identify risk profiles. Furthermore, a claim and risk management outcome would look very different if it was intended to be used as a print-out and handed to risk committee members during a meeting. Therefore, this diversity of application situations illustrates that risk visualization should be used systematically in most risk-related activities.

Comparing the behavior of observations (Q5): In order to investigate how the NLIs can provide the desired visualization responses, in Figure 5.7 I provide the visual response to different questions where different modalities of interactions are utilized. For example, “Show me the distribution of males in the suburb of Lakemba”, “How many females aged 25 are in the state of NSW?” Based on the results, I argue that



1(a). Create a histogram showing distribution of ***State***

1(b). Show me ***Premium-Frequency*** of NSW State

2. Visualize Monthly Benefit of ***Claim-Cause-Code*** in Australia

3(a). Show ***Age*** range between 30 and 70 on X axis and ***Premium*** amount between 500 and 3700 on Y Axis

3(b). Show number of ***Claim-Cause-Code*** from each ***State***

Figure 5.7: Sample questions with answers generated by the system. The answer in Q1(a), and Q1(b) is for query searching, Q2 is for speech and Q3(a) and Q3(b) is for touch/pen/mouse.

the combination of different interaction modalities is a promising research direction in achieving the desired visual response to explore and refine data in an interactive system.

5.5.2 User Study

To observe how visual responses are generated by VAS on the measures of trust and usefulness, I conducted a user study which helps us to understand the penitential utility of the proposed framework. The primary aim of the study is to examine how real users would use the *InsCRMVis* system and to investigate their reaction to multi-modal interaction techniques to explore various queries with several visual views. Thus, the evaluation questions were generated based on the aforementioned key questions (Q1, Q2, Q3, Q4, and Q5) provided in Table 5.4. I performed the study in web-based environments to enhance the system validity, since participants can work on their own.

Table 5.3: User study response for output visualization to post-study questionnaires.

Input	Interaction	User response (with %)	O1	O2	O3	O4	O5	O6	O7
Mouse	Clicking menu	Strongly agree (0.13%)		✓					✓
		Agree (0.27%)	✓		✓		✓		✓
		Neutral (0.27%)		✓		✓		✓	✓
		Disagree (0.07%)				✓			
		Strongly disagree (0.07%)			✓				
Keyboard	Text entry	Strongly agree (0.13%)	✓		✓				
		Agree (0.27%)		✓	✓		✓		✓
		Neutral (0.20%)		✓		✓		✓	
		Disagree (0.13%)		✓				✓	
		Strongly disagree (0.07%)							✓
Speech	Talking	Strongly agree (0.13%)		✓		✓			
		Agree (0.13%)	✓		✓				
		Neutral (0.27%)	✓		✓		✓		✓
		Disagree (0.07%)	✓						
		Strongly disagree (0.07%)				✓			
Touch	Touching menu	Strongly agree (0.13%)		✓		✓			
		Agree (0.20%)		✓		✓			✓
		Neutral (0.20%)			✓	✓		✓	
		Disagree (0.20%)	✓				✓		✓
		Strongly disagree (0.13%)			✓		✓		
Pen	Clicking menu	Strongly agree (0.20%)			✓	✓		✓	
		Agree (0.27%)	✓		✓	✓		✓	
		Neutral (0.33%)	✓	✓		✓	✓		✓
		Disagree (0.13%)			✓		✓		
		Strongly disagree (0.13%)	✓			✓			

Table 5.4: Key observations identified in collaboration with domain experts where expert responses 1, 2, 3, 4, and 5 indicate strongly disagree, disagree, neutral, agree, and strongly agree, respectively.

No	Category	Question	Mean (μ)	Std. Dev (σ)	Min.	Max.
Q1	Interactive	The system let me interact the way I naturally wanted to	2	0.63	1	5
Q2	Insight	I would like to use this system frequently	2	0.63	1	5
Q3	Insight	I found using the combination of mouse, keyboard, speech, touch, and pen to be useful for exploring data visualization	2	0.89	1	5
Q4	Insight	The system enabled me to find interesting insights from the data quickly	2	0.63	1	5
Q5	Speed	I found the answer to my queries about the data	2	0.89	2	5
Q6	Confidence	I found the system was easy to use	2	1.09	1	5
Q7	Confidence	I found the system useful for exploring data visualization	2	1.41	2	5

5.5.2.1 Participants

I performed the user study with 10 participants, 7 men and 3 women, aged between 20 and 59 years. The users were recruited through emailing lists. The participants were mostly students or teachers in universities and stakeholders who had expertise in risk management in the insurance industry. Additionally, participants were familiar with basic data visualizations (e.g., bar charts, line charts, etc.) as they frequently encountered these as part of their study or work.

5.5.2.2 Study Design

To validate the performance of *InsCRMVis*, a TLI questionnaire was administered. A detailed study was conducted to investigate how to visualize information to gain better insights. To visualize the effectiveness of VAS, I implemented five visualization observations such as status position (mouse), text input visualization, speech input visualization, touch, and pen input visualization as shown in Table 5.3. In Table 5.4, the key observation questions were selected to contain satisfactory statements on how the system works and how text/speech/touch output should be structured to avoid distraction. Then, I collected free-form responses as to what the participants considered relevant to the usefulness of the system. The study took about 10 minutes and all the participants worked in automotive research.

5.5.3 Discussion

In this study, the participants rated seven measures on a standard five-point LIkert scale, strongly disagree, disagree, neutral, agree, and strongly agree. The results of these questionnaires are presented in Figure 5.8. I note that the majority of the responses were positive ratings. In particular, most participants agreed that the tool is useful and it enabled them to find interesting insights from the data quickly. More importantly, 6 out of 10 participants found the combination of multiple input modalities to be useful for exploring visualizations.

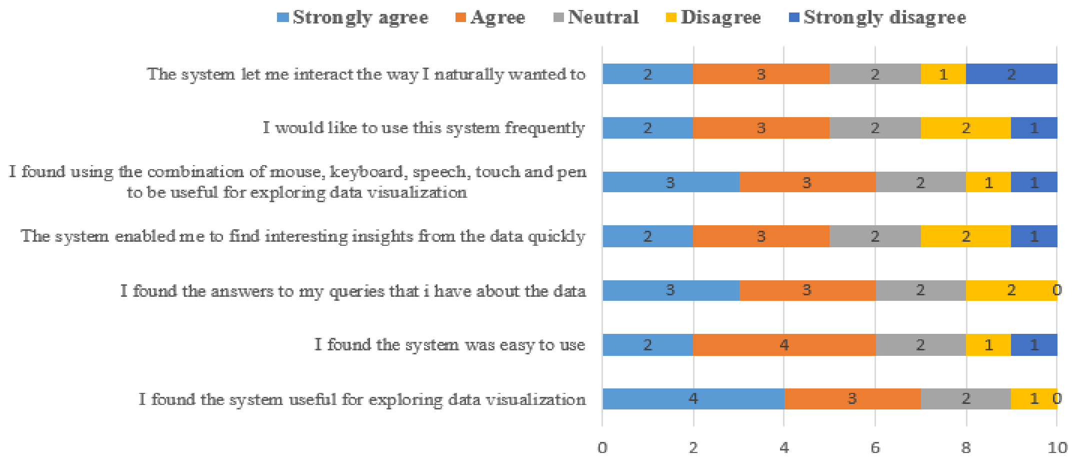


Figure 5.8: Post-study ranking for output visualization and opinion position.

Table 5.3 shows the performance of various input visualization components such as mouse, text, touch etc. Every concept is built on this system. I found the visual responses generated by the system with text queries are more significant than traditional visual outcomes. Additionally, the use of speech and touch were evaluated as both rational and appealing. It is noted that during the conversation with the domain experts about the system status, 75% of the participants positively responded to visualizations of the output text. They informed that the full text was more convenient for utilization than keywords. Additionally, variations in speech output were mostly accepted. Moreover, two-thirds of the 10 suggested suggested multi-modal actions are also appealing to identify visual responses. Thus, the participants' comments were mainly in relation to the user interface, which should be robust, interactive, and smart.

At the end of the study, I provide a series of guidelines that IMs can follow when attempting to visualize risks. The following are the key guidelines:

- Representation of simple text query/conversation can be more flexible to make productive use of visualization in risk management.
- Use of up-to three attributes from the dataset is comparatively more informative to domain experts.
- Use of unnecessary elements in a visualization may cause confusion because of various expectations.
- Various types of risks should be depicted using different queries/symbols.
- Primary risk information should be distinguished from secondary or less important information.

The experts' feedback and user studies ensure the effectiveness of VAS in insurance claims and risk management. I noticed several other challenges that should be addressed. Even though the proposed framework had good outcomes and is valuable for risk visualization, there is room for improvement. For example, if someone wants to get a visual response using a variety of keywords but the proposed framework fails to visualize, this requires the use of transformer (BERT and RoBERTa etc.)-based word embedding methods. The resulting explanation may provide better insights, however, the proposed framework does not provide any explanations for generating a visual response. Another improvement can be to provide guidelines for risk insight visualization.

In summary, I present a web-based visual analytics tool enhanced with query searching, speech, and touch for insurance claims and risk management. I primarily focus on how the system is able to support IMs. I concluded that full text query searching has advantages and provides interesting insights. Additionally, domain experts preferred visualization through speech.

5.6 Summary

In this chapter, I presented a web-based visual analytics solution (*InsCRMVis*) that contains a suite of interactive visualizations, designed in consideration of the task requirements of risk management domain experts. To the best of my knowledge, this is the first work to use natural language interactions with data visualization to address policyholders' claims and manage risk in the insurance industry. In this study, I conducted meetings, interviews, and observational sessions to understand their analysis workflows. The system supports the analysis of multiple types of insurance datasets, such as relational, claim, and demographical. I find that people ask questions and the system provides useful visual insights. The automatic question-answering pipeline achieves an overall accuracy of 69%. I provide a qualitative evaluation of *InsCRMVis* by domain experts based on several use-cases to demonstrate the usefulness of this system in different application scenarios. Additionally, I would like to continue this study with robust methods and more participants in my future work, especially with the spread of age between 20 and 59 years old and different professional groups such as students vs. teachers vs. insurance professionals, etc., as mentioned in section 5.6.

CHAPTER 6

Deep Visual Analytics for Understanding Customer Behavior

This chapter provides the complete development of deep visual analytics (DVA) for understanding customer behavior analysis (CBA). I organized this chapter as follows: First, in Section 6.1, I present the motivation and background study of DVA for understanding CBA. Section 6.2 discusses preliminary work on DVA for understanding CBA followed by the detail methodological discussion in Section 6.3. The description of the visual analytic solutions (VAS) such as *UCBVis* and *Multi-DLMPVis* are presented in Section 6.4. I provide results and discussion of the proposed VAS to assess and discuss their capacity to inform the relevant variables for exploring customer behavior in Section 6.5. Finally, conclusions and future directions are provided in Section 6.6.

6.1 Background and Motivation

We are living in the age of data science (DS), whereas Artificial Intelligence (AI) plays a fundamental role in solving various problems, such as fraud detection, behavior analysis, mental health detection, anomaly detection, natural language processing (NLP) etc. [105]. While most of these tasks seem simple to humans, they are challenging for computer algorithms to solve because there is no systematic explanation. For example, a human can quickly determine whether a picture contains a dog. However, it's difficult to say how they arrived at this conclusion. According to a previous study, various approaches such as visual analytics (VA), data mining, data management, data fusion, machine learning, and other methods have been considered on findings of these solutions [226]. Particularly, VA fosters the practical assessment, correction, and rapid improvement of big data with meaningful interactive visualization (IV).

Nowadays, the use of massive amounts of data is rapidly increasing in many applications. For example, understanding customer behavior (UCB) with multi-dimensional and temporal data is necessary for any competitive global business to provide exciting insights and improve business strategies. The analysis of these data is grimy, irreconcilable, and complex [107]. As a result, vast amounts of time and money are often lost. To address these data issues, new advancement such as VA has proven increasingly efficient and effective in visualizing potential insights in many applications in the past years [102]. For instance, Mandal et al. [179] proposed a novel visual interactive system (VIS) for discovering knowledge and hidden opportunities from massive and complex data. The model automatically learns to bridge the information gap by employing more intelligence in the analysis process and dynamic volumes of information through visual representations and interaction techniques.

Computer-aided many problem-solving techniques have been extensively applied to find individual behaviors, such as dynamic adverse selection, account manipulation, and false invoices [120]. However, they cannot analyze customer sequential claim behaviors, where many customers are involved in claiming their benefits [223]. Moreover, sometimes it is quite complex and challenging to understand, explore, and inspect the individual's potential claim behavioral issues [152, 274]. For instance, (i) the identification of claim behavioral pattern, which depends on the relationship between different customer and their various claim-related attributes; (ii) the uncertainty in the claim history that has arisen in the review procedure; (iii) correlating an extensive amount of claims records and their analysis is time-consuming.

In the existing literature, several researchers explored the traditional machine learning (ML) and visual analytics (VA) techniques by focusing on specific aspects, such as visualization contributes to a better understanding of DL [311], visualization of DL in computer vision [238], visualization for better understanding of ML models [172]; the state-of-the-art predictive VA [175]; interactive machine learning [227]; interpretable ML [170]. For example, Keim et al. [137] analyzed the contrasts between VA and infor-

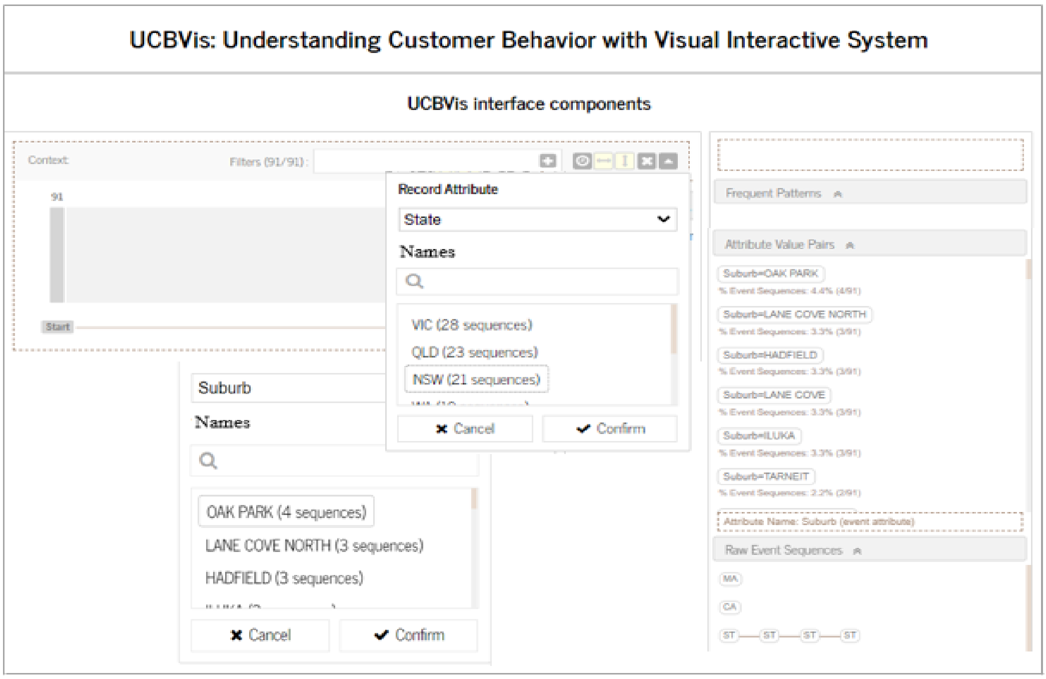


Figure 6.1: Summary of *UCBVis* interface components. (a) behavior exploration workflow: allow user to identify potential behavior of customer, (b) frequent pattern view: allow user to choose sequential patterns for the focus, (c) attribute pair view: allow user to search attribute names with behavioral patterns, and (d) raw sequences view: can be supported with visualization.

mation visualization (InfoVis) from several aspects, including data analysis, perception and cognition, and human-computer interaction (HCI). Caban and Gotz [39] introduced efficient audits of VA approaches which have been proposed to explore complex clinical data. While some of these existing studies have emerged, they are deemed to offer the concept of VA rather than DVA. However, DVA is an advanced development to facilitate visual interfaces, which are flourishing the interactive graphical presentation. Additionally, it plays an essential role in better processing data, utilizing, and visualizing insightful information.

To handle the above challenges, I discuss two-part of work with the insurance collaborator in Australia. The collaborators provided policyholder claim records used in this thesis to understand and explore the detailed requirements of analysts. I examined the related claim data, including policyholder profile, state, suburb, claim date, and claim cause. I combine domain expertise with computing capacity and the expressiveness of visual analytics. First, I present *UCBVis*, an interweaving pattern mining and querying

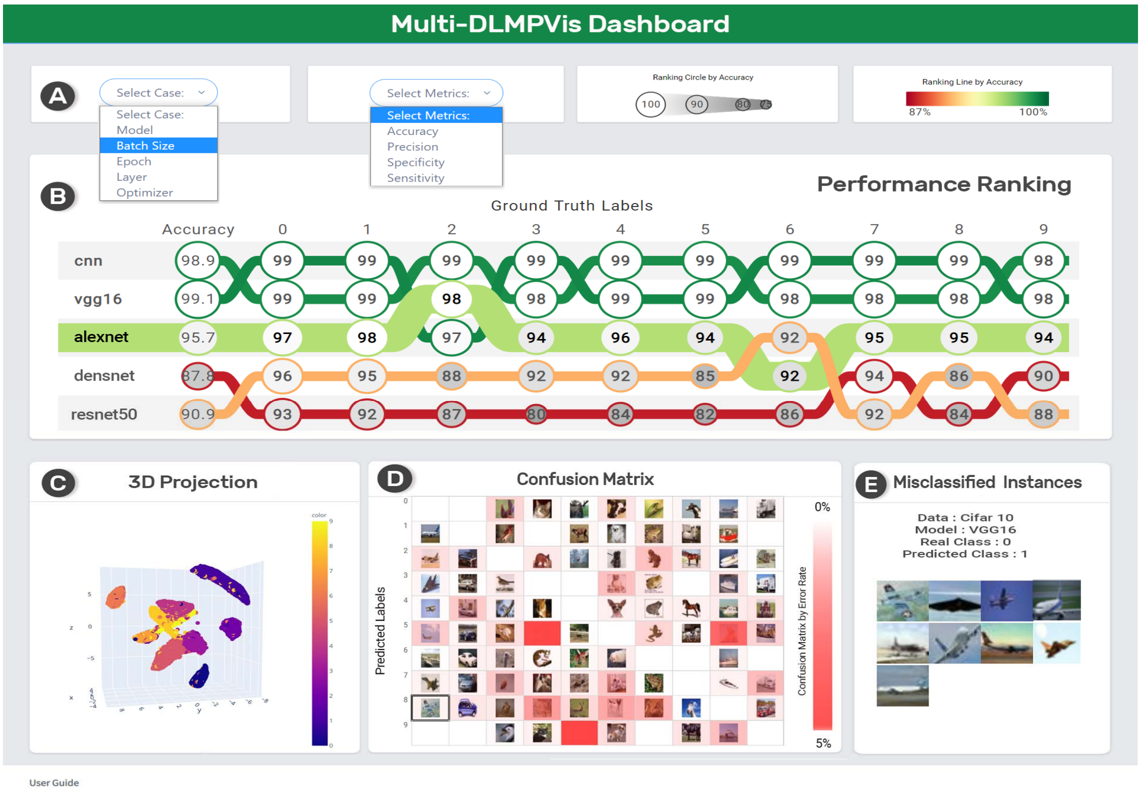


Figure 6.2: *Multi-DLMPVis*: An interactive dashboard for multiple deep learning models performance visualizations (Figure courtesy [8]).

approach as shown in Figure 6.1, to help IMs understand, explore, and inspect the potential claim issues of the customer. I apply *UCBVis* to find a possible activity that was raised in response to the collaborators' requirements. After that, I used this visual design to aid IMs in analyzing customer behavior. Second, according to ahmed et al., [8], I design a VAS named *Multi-DLMPVis* with a wide range of performance evaluation methods that assist the non-expert in adopting an appropriate model, as shown in Figure 6.2. I set the research gap by outlining existing efforts to articulate customer behavior using deep visual analytics. I used a user analysis and expert interviews to demonstrate the efficacy of this method on a real-world dataset. I summarize the contributions as follows:

- I enquire the domain requirements for analyzing customer behavior, together with five domain experts' feedback.
- By interweaving pattern mining and querying with interactive visualization named *UCBVis*, I explore and inspect behavioral sequences from the life insurance claim records.

- I examine tasks of business analysts who aim to understand diverse customer behavior and visualize the outcomes.
- I demonstrate a suite of visualization system named *Multi-DLMPVis* that illustrate the performance of multiple DL models.
- I report a user study with a large dataset and measured professional statements, which show the strength and usefulness of the visualization systems.

In short, I provide a state-of-the-art review on some most significant domains and identify DVA opportunities. I explored various challenges and proposed several directions according to the conducted analysis. This chapter upholds a complete picture of DVA to explore future research by examining the related research in numerous applications.

6.2 Preliminary on Deep Visual Analytics for Understanding Customer Behavior

In this section, I review existing behavioural problems in the insurance industry as well as recent strategies most relevant to my work, such as consumer behaviour detection and deep visualisation techniques and so on.

6.2.1 Exploring Customer Behavior with Visual Interactive System

6.2.1.1 Exploration of Customer Behavior

The understanding of customer behavior is crucial because it has a high demand to the authorities for business management, product marketing, and decision-making [158, 266, 296]. For instances, according to Islam et al. [120], by understanding and visualizing policyholders' claims data, IMs can avoid frauds and to provide risk management in the life insurance industry. Decision makers learn about inter-business activities and can develop new strategies [80]. To monitor a specific stock market user behavior, Huang et al. [89, 112] has provided adverse trading patterns and to identify the real-time stock marker performances. Additionally, to detect user adverse behavior, the coordinated specific keywords visualization is developed within the wire transactions [247]. However, there are very few works on exploring customer claim behavior sequences in the insurance domain. The existing systems have limitation to investigate many variables and satisfy specific requirements, e.g., measuring new claims costs, number of accidental claims, and number of mental health claims of domain experts.

6.2.1.2 Behavior Analysis Techniques

The majority of current literature on customer behaviour analysis is divided into two categories: (i) machine learning and (ii) pattern mining techniques.

Machine learning methods: Machine learning approaches are based on developing models to measure accuracy for a variety of purposes [118, 119, 122]. For example, Amin et al. [10] has claimed that neural networks, decision trees, and Bayesian networks would help in successful evaluation of the impact of suspicious behavior in the business industry. From the findings of their study, it has been determined that suspicious scores of taxpayers are calculated from the information of taxpayers and their related invoice data. Ahmad et al. [7] depict the significance of real-world data for use of modelling as it accounts for more realistic cases. I can see from their research that they propose a hybrid model that combines the support vector machine, multi-layer perception neural network, and logistic regression classification models to detect tax evasion. Through the Wassouf et al. [287] study, I discovered that a multi-class predictive tool was created to detect fraudulent financial misstatements. Hoglund [104] investigates the most influential features in predicting fraudulent tax payment behavior. They state that decision support tool is crucial in reducing fraud tax payment defaults. However, they cannot visualize the identification of insurance claim behavior records.

Pattern Mining Techniques: Pattern mining aid in discovering interesting insights by utilizing various data sequences [51, 173, 183]. For example, Islam et al. [120] propose association rule mining technique to represent the customer adverse behavior. They summarized adverse patterns and presented an unexpected behavioral records to explore suspicious user in the life insurance. Apaolaza and Vigo [13] proposed a hybrid method for identifying suspicious behaviors that using pattern mining model impact performance and that clearance of identifying tax-burden can be evaluated. Graph based components are generally considered having differences between the diversity and accuracy. Thus, Ordonez-Ante et al. [200] investigate an attributed-graph based approach for large scale of claim data is used to detect suspicious records with rule mining technique. Although existing work on combining mining with querying is not a new idea, there has been no systematic investigation into customer behavior exploration in the insurance domain.

Thus, my work lies at the intersection of both lines of research.

6.2.1.3 Data Visualization

To provide possible insights into financial sector applications, I have categorised related financial data visualisation works into three categories: (i) time-series, (ii) multi-attribute, and (iii) financial fraud data visualisation.

1. Time-series Data: To detect, explore, and predict the many specific problems, there have been developed many visual analytic solutions for considering time-series data in diverse fields [81, 163, 261]. For example, a novel visualization system is developed for analyzing spatio-temporal correlations by Malik et al. [178]. Xie et al. [294] introduced a visualisation framework called VAET to distinguish salient transactions from large e-transaction time-series. Yue et al. [309] presented a novel timeline visualisation to look at the evolution of Bitcoin transaction trends from two viewpoints. In addition, existing work on visualising time-series data aims to investigate the interrelationships between them [151, 300]. However, there are some limitations embedded within existing research are required of data visualization for exploration and analysis.

2. Multivariate Data: Multivariate data is employed for the spontaneous and interactive topological data inquiry [127]. To promote global analysis with strong communications to reinforce collections and aggregations, a large-scale multivariate network visualization system is introduced by Dai and Genton [70]. Soriano-Vargas [247] presented simple visual encoding systems to concentrate on exploring the local subgraphs. In another study, they focus on investigating any unusual phenomenon to discover the most suspicious links and to simplify its identification.

3. Financial Fraud Data: The visual analytic system (VAS) has showed its relevance in identifying financial fraud [1, 160]. Existing work by Huang et al. [112] used visualization technique to detect fraud in the financial market. Qu et al. [212] proposed a VA framework to observe real-time stock market activity, control a specific stock, and use 3D treemaps to produce unusual pattern. In another study, Singh and Best [244] demonstrated VISFAN, a visual analytics framework for detecting financial crimes like money laundering and fraud in financial activity networks. EVA is a set procedure with interactive visual analytics facilities proposed by Leite et al. [160], which offers a set procedure with interactive visual analytics facilities to play fraud confirmation.

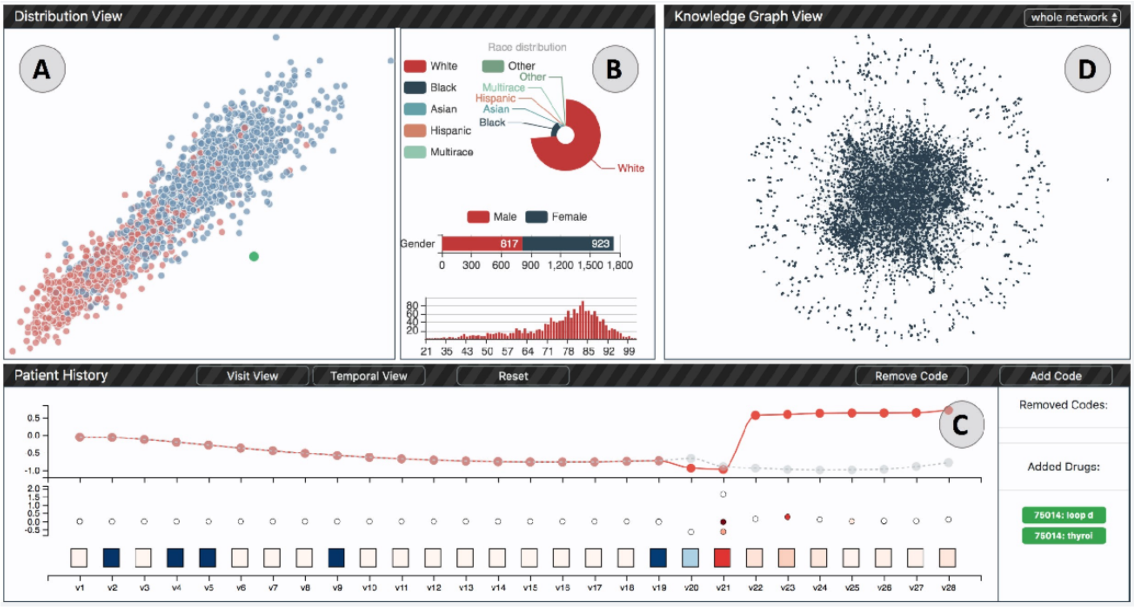


Figure 6.3: An interactive deep visual analytics system which is consisted of four major parts. A) distribution view, B) demographic chart, C) patient history, and D) knowledge graph view (Figure courtesy [166]).

In summary, my research is unique in that it focuses on customer behavior sequences. Thus, to discover customer claim behaviour problems, I presented multiple organised visualisations with carefully developed visual encoding schemes.

6.2.2 Deep Visual Analytics

Visual analytics (VA), a relatively new dimension that has seen rapid growth and considerable interest due to its state-of-the-art success in various sectors, has evolved into deep visual analytics (DVA). While VA attempts to visually represent a dataset with the aims of potentially obtaining some insights, DVA covers the more time-consuming tasks of formulating, refining, and validating theories about the phenomenon underlying the results. Usually, it consists of two major parts such as (i) data visualization which is an emerging field in the current situation [121, 313], (ii) DL which adds more insights, excels at knowledge communication, and discovering strategies by applying encoding techniques to transfer abstract data into meaningful representation [79]. Figure 6.3 shows an interactive clinical prediction visualization system, whereas DL brings an extra power to predict clinical risks. Hence, DVA's ability to change visualisations interactively and pose complex on-the-fly identified queries is crucial.

Table 6.1: Overview of key representative works in visual and deep visual analytics.

Year	References	Big data analysis	Cognitive and perception science	Customer behavior analysis	Natural language processing	Recommended system	Healthcare analysis	Fintech ecosystem	Tourism management	Publication Venue
2021	Mandal et al. [179]	✓								DCN
	Krak et al. [145]		✓							RASTORE
	Wang et al. [286]				✓					arXiv
2020	Chang et al. [46]		✓						✓	TM
	Li et al. [166]						✓			JMIR
	Mubarak et al. [188]	✓								WOL
	Qian et al. [210]					✓				arXiv
	Shin et al. [243]	✓								MIS
	Wang et al. [283]	✓								APVIS
	Zeng et al. [312]	✓								IEEE VCG
	A. Leite et al. [1]			✓						CGF
2019	Vellido et al. [273]						✓			NCA
	Ku et al. [147]								✓	HICSS
	Aupetit et al. [19]		✓							IEEE VIS
	Zhang et al. [314]		✓						✓	TM
	Hu et al. [110]					✓				HFCS
	Hohman et al. [106]	✓								IEEE VCG
	Hohman et al. [106]				✓					IEEE VCG
	Yue [308]							✓		PhD diss
	Singh and Best [244]			✓						IJAIS
2018	Kwon et al. [150]						✓			IEEE VIS
	Cashman et al. [41]	✓								IEEE CGA
	Garcia et al. [91]	✓								CG
	Peixinho et al. [204]				✓					CGIP
	Choo and Liu [55]		✓							IEEE CGA
	Yue et al. [309]							✓		IEEE VCG
2017	Liu et al. [171]			✓						IEEE ITS
	Pezzotti et al. [206]	✓								IEEE VCG
	Samek et al. [231]	✓	✓		✓					arXiv
	Leite et al. [160]			✓						IEEE VCG

6.2.2.1 Why Deep Learning for Visual Analytics

There are various domains where VA has been proposed to help model developers to build, debug, and precipitate the experimental process to improve performances [148, 255, 291]. For example, analyzing medical images [277], explaining decisions made by medical imaging models [315], feature extraction from imaginary instances which are helpful for designing and planning [320]. However, it is required to monitor during the training phase, distinguishing mis-sorted instances and testing the well-known data instances to improve performance [43, 221]. Therefore, the advanced development of DL based VA provides a better solutions to enhance the model development process for engineers and researchers, improve overall performance, accuracy and speed up the debugging.

Deep visual analytics (DVA) is an extended part of the visualization field. It can rather be seen as an integral approach to make decision, combining visualization with DL models, human factors and data analysis [122, 304]. Nowadays, DL models are most useful for decision making as well as providing as many accurate predictions as possible [105]. It is persistently using for decision-making tasks [195]. The most significant reason behind DVA is to provide a better understanding of the properties of the input data and visualizing their demonstration [230]. To more readily work with interpretability and explainability [170, 187], interactive DVA solutions have been proposed to help various user groups interpret models using an IV [2, 106]. Although some research works have explored DL in VA, there are enough space to resolve various problems in the existing works. Table 6.1 present various key representative works in visual and deep visual analytics. From existing research works, it has been cleared that DVA are much more interpretable, explainable and scalable for decision making, comparing model performance, debugging, and feature extracting. DVA focuses on the experimental process rather than theoretical explanation. Moreover, with the constant advent of novel research works, it has been perspicuous that DVA are much more comprehensive and suitable and a new inclusive framework for better understanding of these dimensions.

6.2.2.2 Impact of Deep Visual Analytics for Understanding Customer Behavior

With the advanced visualization modules, such as deep visual analytics (DVA), the visual analytics (VA) researchers have developed intuitive and immersive user interfaces. Such

DVA systems offer users a thorough understanding of a model and hints on how to troubleshoot and develop it. In the recent years, several studies have been conducted, where DVA has proven their effective and noteworthy impact in many fields, especially in big data analysis [243], human cognitive and perception science [314], analyze healthcare systems [166], and tourism management system [46]. In addition, DVA has the greatest impact of synthesizing information, deriving insights from massive, ambiguous, unstructured, unexpected patterns, decision making, and communicating assessment for action [137]. For example, CNNVis [172] is a good example of a VA framework for understanding CNN models and diagnosing them. ActiVis [131] uses several organised views, such as a matrix view and an embedding view, to provide a visual exploratory overview of a given DL model. ReVACNN [60] provides realtime model steering capabilities during training, and interactively selecting data items for a subsequent mini-batch in the training process. Besides, DVA was investigated in the medical sector by Li et al. [166], which improved model efficiency and paved the way for interactive, interpretable, and reliable clinical risk predictions.

Overall, DVA's impact is determined by its ability to include advanced visualisation and interaction capabilities. However, several research issues such as how to efficiently loop humans into the analysis process and how to increase the applicability of DL techniques have not been thoroughly explored. Hence, research has shown that DVA is more effective, impactful, and most of the DVA tools attract more attention to validate the performance of deep models as well as rapidly spreading their effectiveness on multivariate sectors and also provide visual interactivity to DL experts.

6.2.3 Strategies for Evaluating Visualization System

Nowadays, various complex problems have been explored by visual analytics systems (VAS) globally where a proper evaluation provides the necessary insights into a visualization system that enhances VIS more profoundly. The importance of evaluating the visualization system has become well-recognized and demonstrated by the growing body of work on how to conduct visualization evaluation and the increasing amount of research papers that incorporate a formal or informal assessment. For example, Mandal et al. [179] proposed a novel VIS for discovering knowledge and hidden opportunities from massive and complex data. The model automatically learns to bridge the information gap by employing more intelligence in the analysis process and dynamic volumes of information through visual representations and interaction techniques.

I provide a systematic assessment and understanding of the evaluation practices reflected by peer-reviewed journals and conferences. I studied several evaluation methods such as the Likert scale, eye trackers, log data analysis, comparing dashboards, insight-based evaluation, qualitative and quantitative feedback analysis, long-term case analysis, and Nielsen heuristics to select a standard evaluation method. These evaluation methods have been applied more to evaluate any visual analytics systems, which has been declared in table 6.2. Moreover, research development trends show that these evaluation methods have increased recently. In short, I have reviewed more than a hundred articles and mostly applied these evaluation methods for visual analytics systems.

6.3 Methodology

This section describes two types of design methodology such as *UCBVis*, and *Multi-DLMPVis* are employed as a designed dashboard to bridge the information gap by using more intelligence in the analysis process and dynamic volumes of information through visual representations and interaction techniques.

6.3.1 Methodology of *UCBVis*

This section introduces the design process by describing the data, domain goals, and requirements that experts hope to accomplish with the visualisation framework.

6.3.1.1 Data Description

In this study, I collected a large scale dataset from one of the local Australian life insurance companies to study customer behaviour analysis. The dataset consists of claims record in over ten years periods (2010 to 2019) from 13,287 countrywide policyholders. Each record comprises four attributes: ID (individuals unique ID), Claim-cause (what is the claiming reasons), State/Suburb (where they claim is recorded), Date (when and how frequently they are claimed). The detail of the dataset is described in Table 6.3.

6.3.1.2 Domain Goals and Requirement Analysis

To better understand the customer behaviour and explore insights, I contacted with domain experts, who have been working over the past five years in the insurance industry. The experts have solid experience, and they are proficient for claim risk management in

Table 6.2: Overview of various evaluation techniques and their applications.

Year	References	Likert Scale	Eye trackers	Log data analysis	Comparing dashboard	Insight based evaluation	Qualitative Analysis	Quantitative Analysis	Nielsen heuristics	Publication Venue
2021	Zerafa et al., (2021) [313]				✓					IEEE ICIV
	Qian et al., (2021) [211]							✓		ACM
2020	Sahu et al., (2020) [228]							✓		IEEE CCWC
	Qian et al., (2020) [210]	✓								arXiv
	Kandasamy et al., (2020) [134]	✓								MA
	Stehle et al., (2020) [252]				✓					IJGIS
	Samuel et al., (2020) [232]						✓			MDPI
	Li et al., (2020) [166]				✓		✓	✓		JMIR
	Wang et al., (2020) [283]						✓	✓		IEEE PVS
	Beasley et al., (2020) [27]	✓								PacificVis
2019	Yu et al., (2019) [305]			✓						IEEE TV
	Lee et al.,(2019) [157]						✓	✓		IEEE VCG
	Bourqui et al., (2019) [36]				✓					arXiv
	Ku et al.,(2019) [147]						✓	✓		HICSS
	Steyn et al., (2019) [253]						✓			AEHE
	Haleem et al., (2019) [97]							✓		IEEE CGA
	Chen et al., (2019) [48]		✓							ICHCI
	Dibia et al.,(2019) [76]				✓					IEEE CGA
	Hu et al.,(2019) [110]						✓	✓		ACM CHI
	Deng et al.,(2019) [74]						✓	✓		IEEE CGA
2018	Yasmin et al., (2018) [299]	✓								IEEE ICNOF
	Garcia et al., (2018) [92]						✓			CGF
	Saggi et al., (2018) [48]					✓				IPM
	Barcellos et al., (2018) [25]								✓	IEEE ICIV
	Wang et al., (2018) [281]					✓				IEEE CGA
	Kwon et al., (2018) [150]						✓	✓		IEEE CGA
	Chang et al., (2018) [44]		✓							PacificVis
2017	Shao et al., (2017) [241]		✓							ACM
	Ming et al., (2017) [185]					✓	✓			IEEE VAST
	Swaid et al., (2017) [258]								✓	ICHCI
	Kahng et al., (2017) [131]					✓	✓			IEEE CGA
2014	Blascheck et al., (2014) [32]		✓							EuroVis

Table 6.3: Demographic characteristics of claim data set.

Variables	**Category**	**Number of policyholders**	**Percentage (%)**
Policy ID	Number	13287	100%
Claim-cause	Depression	326	2.45%
	Stroke	316	2.37%
	Neurotic disorder	1891	14.23%
	Accidental Falls	6192	46.60%
	Motor Vehicle Traffic Accidents	777	5.85%
	Cancer	2685	20.2%
	Unemployment	1364	10.26%
State	NSW	3566	26.83%
	VIC	2918	21.96%
	QLD	4099	30.85%
	SA	999	7.52%
	WA	1394	10.49%
	TAS	311	2.34%
Suburb	Whole Australia		
Date	2010-2019		

Australia. Face-to-face interviews and private meetings with domain experts were also held. Therefore, I am interested in three aspects of customer behaviour analysis:

- What is the standard practice and procedure for identifying and analyzing customer behaviours?
- What are the main challenges and limitations of the ongoing approaches for detecting and analyzing customer claim behaviours?
- What kind of study concerns and tasks do they prefer to bring?

Through the interview sessions, the experts analyzed the current analysis strategy, which comprises three steps:

- Customers' period-end claim positions.
- Require new fieldwork and proper business records.
- Analysis of insightful information requires a vast amount of manual controlling of the raw claim data, which is tedious and time-consuming.

The above design requirements are shown in Table 6.4.

Table 6.4: Domain specific requirements.

SN	Requirement
R1	The system should allow analyses to be performed over selected suburbs of the city (the whole city or a part of it, e.g., Redfern in Sydney).
R2	The system should allow the types of claim-cause to be analyzed individually.
R3	The system should be simple and intuitive, allowing any officer to operate it with minimal training and basic computer skills.
R4	The system should allow for annotations and sharing, also allowing for continuous and shareable analyses.

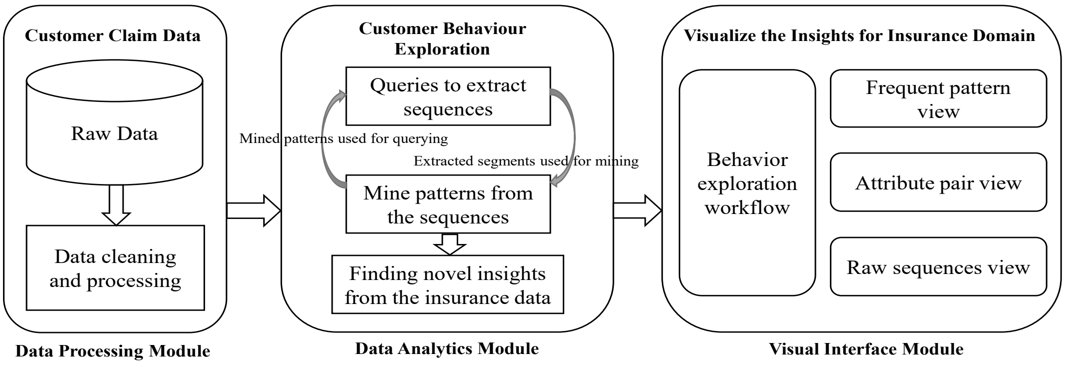

Customer Claim Data
Raw Data
Data cleaning and processing
Data Processing Module
Customer Behaviour Exploration
Queries to extract sequences
Mined patterns used for querying
Extracted segments used for mining
Mine patterns from the sequences
Finding novel insights from the insurance data
Data Analytics Module
Visualize the Insights for Insurance Domain
Behavior exploration workflow
Frequent pattern view
Attribute pair view
Raw sequences view
Visual Interface Module

Figure 6.4: The architecture of *UCBVis* system.

6.3.1.3 Implementation Process

The study is not about designing a new visual analytics system (VAS), rather considering an existing interactive VAS [154], I play a fundamental role in supporting the analysis task of business domains which requires thorough task analysis and domain expertise. I customize the design framework named *UCBVis* of an existing system for analyzing customer behaviour and to help the insurance risk manager. Thus, by following the study requirements discussed in Table 6.4, *UCBVis* first identifies customer behavior sequences and then visualizes the outcome. The system integrates pattern mining and querying techniques with an interactive visual interface. As shown in Figure 6.4, *UCBVis* is made up of three main modules: (i) data preprocessing, (ii) data analysis, and (iii) Visual Analysis.

The data processing module performs both data masking and model construction. As mentioned in Subsection 6.3.1.1, I record various information in the dataset with comprising various attributes. Since attributes have values in different categories, it may contain missing values. To simplify the system to ensure I use only the most significant data, data preprocessing comprised reducing less important and redundant attributes that offer no benefit to exploration and analysis. As part of data preprocessing, redundant fields that were not eliminated were combined. The data analysis module interweaves the mining and querying based on the parameters used to find novel insights from the insurance data. I allow analysts to evaluate behaviour sequences based on the exploration goals. To make it easier for people to explore themselves, I extract potential patterns from the data, and these patterns characterize individual behaviours of claim issues. Via intuitive visualisations, the visual analysis module introduces customers' different behavioural patterns, as well as the underlying contexts and comprehensive proof. It simplifies an accessible analysis and in-depth search of individuals with these related visualization views. I have presented these modules as a web-based system.

The *UCBVis* system is developed as a web-based application, where Python is used to develop the backend to support data processing and analysis. JavaScript is used to implement the frontend where Data-Driven Documents (D3) is used to build visualization views. The interface is made up of HTML/Scalable Vector Graphics (SVG) components, and the web application is structured using the AngularJS framework, which follows the model-view-controller paradigm.

6.3.2 Methodology of *Multi-DLMPVis*

According to Ahmed et al., [8], this section discusses how *Multi-DLMPVis* is used as a designed dashboard for examining visual performances of multiple deep learning models and the details of data gathering and processing.

6.3.2.1 Data Description

The dataset I applied was the CIFAR-10, which is freely available [146]. It is one of the most popular datasets for deep learning and image processing research. It contains 60,000 32x32 color images divided into ten categories, each with 6,000 images: automobile, airplane, bird, deer, cat, dog, horse, frog, ship, and truck.

6.3.2.2 Design Consideration

As demonstrated in Figure 6.2, I designed an VAS named *Multi-DLMPVis* to explain and visualize several DL model performances. I go over how it works with image data and provide different metrics and model results. It also demonstrates the classification

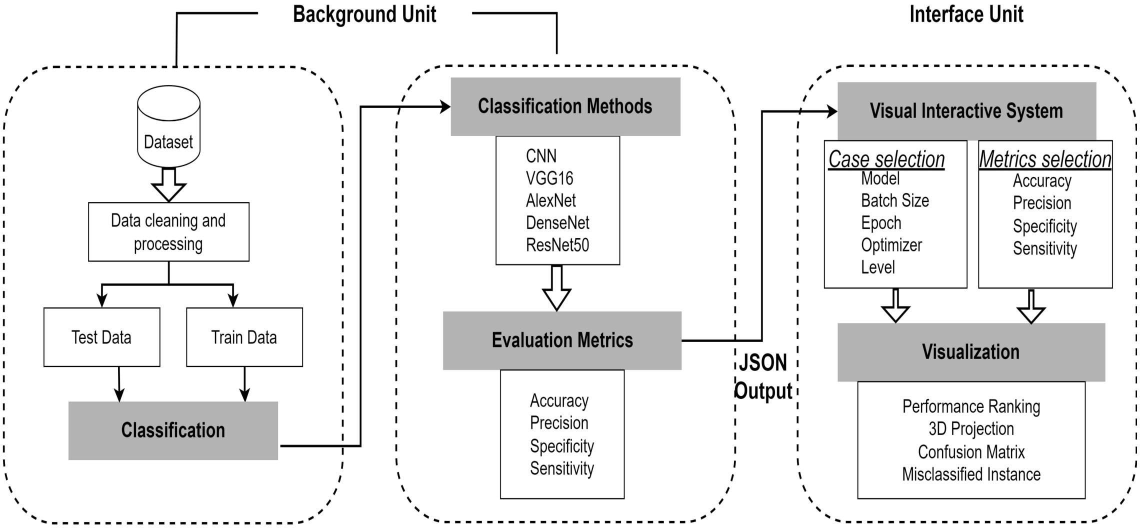


Figure 6.5: System architecture of *Multi-DLMPVis* System.

performance in 3D projection with misclassified instances and shows multiple evaluation metrics outputs. Figure 6.5 shows the system architecture of Multi-DLMPVis, which comprises two modules: (i) background unit, and (ii) interface unit.

Background Unit: The background unit of *Multi-DLMPVis* serves as a data processor. First, it accepts a dataset as an input, prepossess, then divides it into training and testing subsets. The dataset is trained on multiple DL models and calculates high-level information about each model's performance to create the performance result summary. For example, model name and hyperparameter include layer structure, epoch number, batch size, optimizer, and metrics in the model's performance. Finally, the performance measurements for the model are saved in a JSON file, and the data is sent to the interface unit for displaying.

Interface Unit: The interface unit of *Multi-DLMPVis* comprises five modules, as shown in Figure 6.2: a) Allowing users to select hyperparameters and metrics; b) A bump chart performance ranking view at the class level displays many models' performance; c) A 3D projection of the latent space represents a hidden feature in each model; d) A view of the confusion matrix to check each model's flaws; e) This module shows cases that have been classified. Using this system, users may quickly evaluate several deep learning models and gain actionable insights for creating the models or deciding which model to adopt.

6.4 Experimental Analysis

6.4.1 Experimental Analysis of *UCBVis*

As shown in Figure 6.4, *UCBVis* comprises four major views: (i) Behavior extraction workplace, (ii) the frequent pattern view, (iii) the attribute-value pair view, and (iv) raw sequence view. In this section, I describe how each workplace supports the user analytic tasks.

6.4.1.1 Behavior exploration workplace

The "Behavior exploration workflow" as shown in Figure 6.4 helps analysts who do not have an exact idea about how a VAS is being used to showing the full dataset as the context. It comprises several functionalities such as top bar, flow visualization, timeline. In order to promote the exploratory analysis, analysts can select "state" information, choose various "suburb area" that remain in the dataset to the flow visualization.

6.4.1.2 Incorporating frequent pattern into the workflows

As illustrated in Figure 6.4, the "Frequent Pattern View" component provides diverse information that is recorded various attributes and applied as input for the pattern mining algorithms. There are two different forms to get behavioral information from the frequent pattern workflow. First, users can use "Query Inputs" of the executed queries from the interface to take as input the results. Second, the system obtains a set of attribute under the corresponding header in the pattern analysis component, which includes the claim-cause and users' other information discussed earlier. Once the corresponding inputs are selected, users can get the raw interaction data through the "Attribute name selection" functionality. Additionally, user can set the minimum support parameters.

6.4.1.3 The attribute pair view

The extracted values are shown as a list in descending order of the level of sequences in the centre that contain claim records in the "attribute pair view." The list is made up of attribute pairs that all belong to the same attribute. By tapping on the attribute name, analysts may change the characteristic. When analysts hover over a claim attribute pair, a button appears, allowing them to break the emphasis by that attribute pair. In addition, when you click this button, a new panel will appear with the focus separated and visualised.

6.4.1.4 Raw sequence view

In the “Raw sequence view”, analysts can extend the chances of locating relevant patterns by customising the behavior set to be applied as input for pattern mining. I can utilize the behavior selection to adjust the input, putting on specific sequences of behaviors. Additionally, custom behaviors set up using the “Query Search” view by any user of the system, can also be added in the behavior set.

6.4.2 Experimental Analysis of *Multi-DLMPVis*

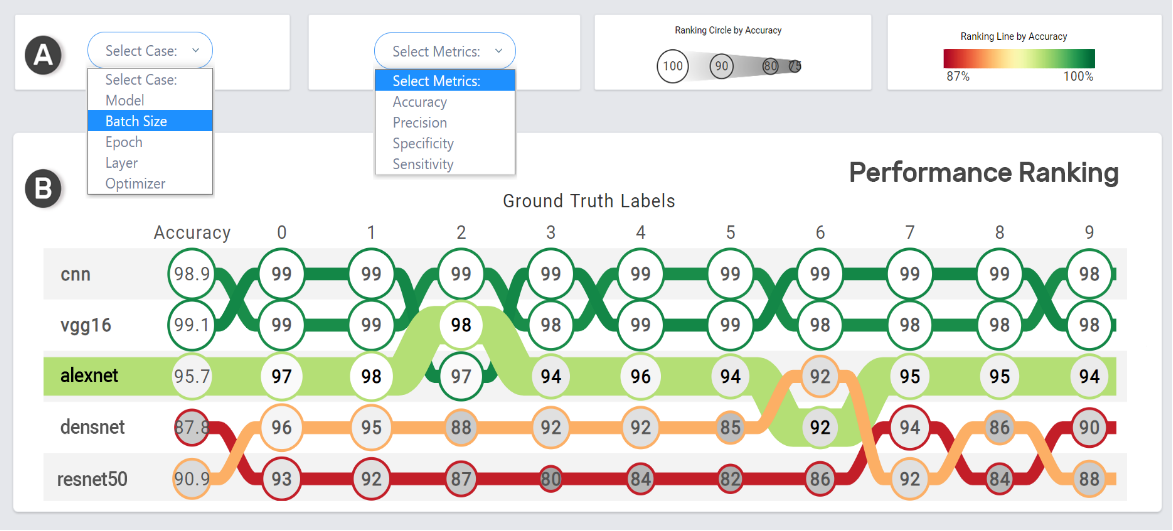
This section presents how *Multi-DLMPVis* examines images within different DL models, such as CNN, AlexNet, VGG-16, ResNet-50, and DenseNet perform. The dashboard of the *Multi-DLMPVis*, as illustrated in Figure 6.2, is made up of five major modules as follows:

6.4.2.1 Input Selection

The *Multi-DLMPVis* input selection allows users to filter hyperparameters like batch size, epoch, layer, and accuracy metrics. As demonstrated in Figure 6.6, they can interactively explore and compare models based on performance rankings and ground-truth labels.

6.4.2.2 Performance Ranking

The performance rating module allows DL researchers to compare and contrast the results of numerous models simultaneously. They can evaluate any class’s overall performance and individual performance metrics. I created a bump chart to visualize multiclass and multimodel performance simultaneously, as illustrated in Figure 6.6. Each model is represented in the graph by a ranking line, with columns representing the ground truth class level. Each class is represented by a circle, which contains the performance measure value. The size of the circle changes depending on the class’s measured value. The red to the green color scheme in the ranking line corresponds to the performance measure value of the models. When a user selects a model on the chart, the ranking line for that model becomes bold, making it easy to see which class is performing poorly.



	Accuracy	0	1	2	3	4	5	6	7	8	9
cnn	98.9	99	99	99	99	99	99	99	99	99	98
vgg16	99.1	99	99	98	98	99	99	98	98	98	98
alexnet	95.7	97	98	97	94	96	94	92	95	95	94
densenet	87.8	96	95	88	92	92	85	92	94	86	90
resnet50	90.9	93	92	87	80	84	82	86	92	84	88

Figure 6.6: The input selection and performance ranking view of the *Multi-DLMPVis* System.

6.4.2.3 3D Projection

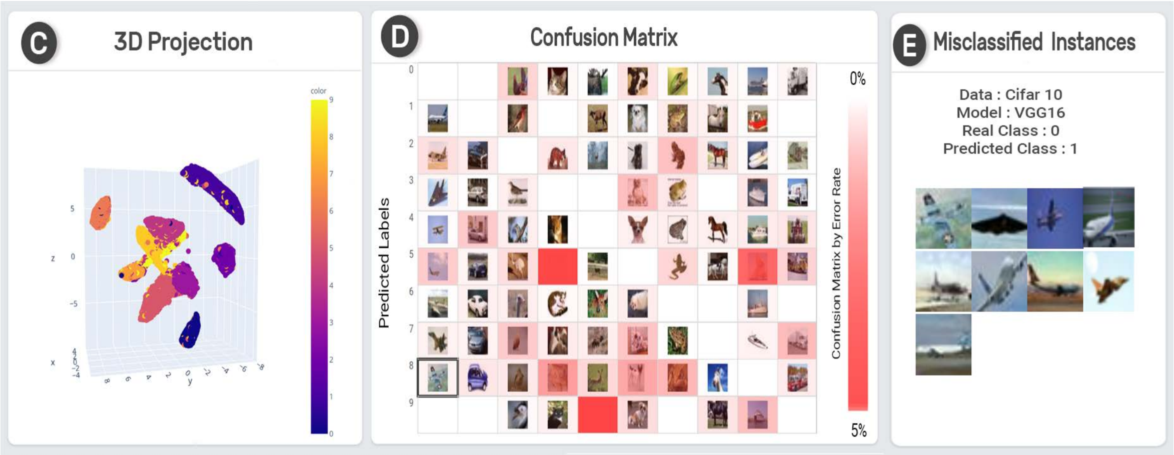
The latent space learned by the model is presented in the *Multi-DLMPVis* 3D projection. Each point in Figure 6.7 represents a single instance, and the image's ground truth label determines the color of the point. The t-SNE input is the activation values from the last hidden layer of the DL models, and the t-SNE output is the presented 3D scatterplot [271].

6.4.2.4 Misclassified Instances

Multi-DLMPVis allows DL practitioners to identify misclassification tendencies quickly. Practitioners can spot patterns in mislabelled cases and look into why they were mislabelled in the first place. As shown in Figure 6.7, all misclassified cases are based on the model confidence of the misclassification class.

6.4.2.5 Confusion Matrix

The confusion matrix view displays the anticipated instances of class and actual data classes, respectively, represented by the values in the row and column as shown in Figure 6.7. Each cell contains an image of a misclassified instance for a better understanding. For example, a red color density marks misclassified cases. The higher the color density



Data : Cifar 10
Model : VGG16
Real Class : 0
Predicted Class : 1

Figure 6.7: 3D projection, confusion matrix, and misclassified instances view of the *Multi-DLMPVis* System.

on each cell, the greater the possibility of misclassification. In short, this system allows users to immediately detect the model’s flaws at the level of class and confusion.

6.5 Discussion

6.5.1 Evaluation

I conduct a two-stage evaluation study to assess the potential usability and usefulness of the system. In the first stage, I created a set of questions as shown in Table 6.5 and asked five (5) participants to freely use the system and provide feedback about the usability and utility. Each participant were used individually and provided feedback. In the second stage, *Multi-DLMPVis* is compared with three different interactive visualization dashboards as shown in Table 6.6 where (✓) and (x) indicates the presence and absence of the feature selections [52], [131], [282]. In short, the system is useful for evaluating performance at the class and instance levels to compare multiple models effectively.

6.5.2 Challenges

Visual analytics (VA) is an application oriented discipline where DL techniques have built significant advancements and its research venues being in the limelight. From the existing studies, it is observed that with the acute development of DVA, many complex problems have been solved in the application of perception and cognitive science, information management, tourism sector, statistical analysis, knowledge discovery, financial

Table 6.5: User study results.

No	Category	Question	Mean (μ)	Std. Dev (σ)	Min	Max
Q1	Easy to use	*UCBVis* and *Multi-DLMPVis* was easy to learn and use.	3.8	0.75	3	5
Q2	Insight	*UCBVis* and *Multi-DLMPVis* was useful to explore insights patterns.	3.6	0.80	3	5
Q3	Insight	*UCBVis* and *Multi-DLMPVis* allowed me to discover insightful queries about the data.	3	0.89	2	4
Q4	Essence	*UCBVis* and *Multi-DLMPVis* helped me to generate knowledge about the claim data.	3.6	1.02	2	5
Q5	Speed	*UCBVis* and *Multi-DLMPVis* enabled me to find interesting insights from the data quickly.	3.2	1.16	2	5
Q6	Confidence	*UCBVis* helped me to grow confidence about the interesting data insights.	3.5	0.57	3	5
Q7	Confidence	*Multi-DLMPVis* helped me to grow confidence about the interesting data insights.	4	0.63	3	5

analysis and medical sector. Additionally, DVA can be assessed based on their final outputs without the understanding of how they get to these decisions. However, several application domains are practically untouched by DVA because of their challenging nature or the lack of data availability. Therefore, in the following, I outline key insights into their challenges for doing future research using DVA.

1. Data scalability: Scalability is an important aspect that I considered during the development of *UCBVis and Multi-DLMPVis*. For example, due to the limited claim records, *UCBVis* works well for customer behavior analysis. However, when there are significantly more claim records, it may be challenging to show the result in visual clutter overview. In terms of visualization, an improvement could be displayed.

2. Data availability and design choice. The system relies on expert feedbacks, where they stated that the policy premium of the customer plays a significant role in claim inspection. However, in this study, the policy premium is not available in the

Table 6.6: Feature comparison analysis.

No.	Features of dashboard	ActiVis [131]	DeepVID [282]	DECE [202]	***Multi-DLMPVis***
F1	Multiple Model Compare	x	x	x	✓
F2	Performance ranking	x	x	x	✓
F3	Performance metrics	✓	✓	✓	✓
F4	3D projection mapping	x	x	x	✓
F5	Hyperparameter Tuning	x	✓	x	✓

Notes: (✓) indicates the presence of feature selections and (x) indicates that it does not visualize the feature outcome.

dataset. Once the policy premium information is accessible, the system can provide more concrete evidence. For example, the system could visualize each policyholder's location where they are living so that IMs can monitor which suburbs are risky.

3. Human computer interactions: According to current research, developing an interactive visual interface is important that reduces the gap between the human's cognitive model of what they want to achieve and the computer's understanding of the human's task.

4. Evaluation: Human information discourse constitutes a challenge for evaluating DVA applications' utility, effectiveness, and trustworthiness. Data uncertainty may arise during the analysis process, which misleads decision-making and analysis results.

5. Unstructured and unlabeled data: Unstructured and unlabeled data from heterogeneous sources reduce the accuracy, cause data loss and generate wrong patterns.

6. Unexpected pattern: Uncertain and misleading data generates unexpected patterns that diminish the outcome's accuracy.

7. Visualization designs and usability: The visual interface of *UCBVis and Multi-DLMPVis* is simple and easy to understand. For instance, the overview of the behavioral patterns employs the visual design of claim records. These visualization designs are straightforward. However, a VAS with better measured intuitive forms will be significantly simpler to find out the usability of *UCBVis and Multi-DLMPVis*.

6.5.3 Future Directions

In this article, various existing key efforts have been carried out related to the use of DVA from different perspectives. However, there are still enough spaces that need to be discussed. Therefore, several potential future research directions are summarised as follows:

1. Explainable visualization system: Explainable visualisation has pushed the state-of-the-art in deep learning to new heights, and humans now rely on explainable visualisation techniques more than ever before. DVA has had a significant influence on a number of long-standing issues, such as computer vision, speech recognition and synthesis, and NLP. As humans rely on explainable VA, it will be able to interpret their decisions and control over their internal processes for various high-impact tasks.

2. Adverse behavior identification: The process of understanding and monitoring with the help of interactive visualization is very important for business authority to solve real-world problem. Many researchers have applied pattern mining techniques in diverse sectors and achieved the expected outcome [120]. However, with the advanced development of DVA, exploring and visualizing adverse behavior will give more adequacy, especially in data analysis. To advance pattern exploration, I need to involve advanced technique for exploring adverse behavior of users. Thus, DVA can explore the erroneous behavior, particularly in tourism and customer behavior analysis sector.

3. Visual sentiment analysis: Sentiments are emotions and feelings that are sometimes expressed through opinions, likes & dislikes, and symbols. It can be expressed through text, images, audio, and videos. Several researches have done for analysis of sentiments, however, not much work is carried out pertaining to visual sentiment analysis. Additionally, lot of traditional techniques such as CNN, RNN, SVM, RF, PCA etc. are applied for solving various issues and challenges encountered in sentiment analysis, very few works have done using DVA. Thus, advanced development of DVA has numerous advantages which can play a significant role in analysing sentiments from visual data.

4. Risk management: Overseeing and communicating risks have become crucial tasks, when analysing numerous data-sets. Visualizing several risk factors could help to accurately predict data and control the cascade of false data. Moreover, DVA has demonstrated to be very effective and promising, which can play a significant role in assessing risks and identify fraud activities. With the advanced development of DVA, I would be able to contemplate the impact of various risks, interventions and correction techniques on a large scale, to better understand their impact on numerous sectors.

5. Multi-task learning: In many DL tasks, from computer vision to NLP, multi-task learning have seen very good progress. In recent years, several researchers have applied multi-task learning to visualise data with DN framework and found that it outperforms over single task learning. The benefits of using DNN-based multi-task learning are threefold: (i) learning several tasks at once prevents overfitting by generalising hidden representations; (ii) auxiliary task provides interpretable performance for explaining the visualization outcomes; (iii) multi-task provides implicit data augmentation to alleviate the sparsity problem. Thus, I can use multitask learning for cross domain recommendation in addition to adding side tasks.

6.6 Summary

In this chapter, I presented *UCBVis, and Multi-DLMPVis*, an interactive visualization system for any competitive and global business aimming to provide interesting insights and to improve business strategies. Using *UCBVis*, I present customer behavior pattern of multiple relationship through the visualization system based on interweaving the pattern and querying with a designed encoding scheme. Second, the findings of *Multi-DLMPVis* are visualized into five parts: input selection, performance rating, 3D projection, confusion matrix, and misclassified instances. It also shows the misclassification instances of the confusion level, with several evaluation criteria such as accuracy, precision, sensitivity, and specificity.

In summary, by reviewing related research in different application, this chapter has drawn a complete figure of DVA in order to coordinate future exploration. The state-of-the-art DL techniques and implementations of VIS in various application can solve any issues between the challenges of discovering information in broad and complex data sets.

CHAPTER 7

Conclusion and Future Work

7.1 Contributions

This thesis provides a broader and comprehensive overview of quantitative research for analyzing customer behavior in the life insurance industry. Data mining provides several popular approaches for analyzing customer behavior, which is crucial for business planning and decision-making. Existing methods cannot accommodate various situations due to the growing desire for exploring deeper insight into customer behavior. This thesis investigated and developed innovative data mining approaches blending with interactive visualization systems for analyzing customer behavior using the insurance dataset as a testbed. Therefore, the research was carried out the following aspects:

1. The motivation and several challenges of the existing research have illustrated as essential issues of customer behavior analysis (CBA) in an insurance perspective. Several objectives followed by the proposed methods of decision-making modeling with limitations have been discussed to address these problems. Finally, the organization of this thesis is outlined (Chapter 1).

2. Summarize a detailed description of the understanding of CBA, visual analytics in CBA, and data mining techniques for business risk management and identify several potential research issues in the insurance domain (Chapter 2).

3. Customer decision-making models that are effective should assess numerous factors simultaneously and account for all possible interactions between them. Traditional techniques are inadequate to meet these complex needs. There was no meaningful insight into the decision-making process. Thus, this thesis proposes a pattern mining technique for evaluating policyholders' questionnaire data and then researching policyholders' adverse behavior to address the problem of decision-making modeling (Chapter 3).

4. Exploring customer behavior has presented a new dimension for studying customer behavior with great potential. Unfortunately, no standard framework has been available to visualize such data effectively, and limited tools support the analysis process and application development based on these data resources. Aiming to address this limitation, this thesis focuses on designing an interactive visualization system to explore profound insights into customer behavior and the complex decision-making process of customers based on pattern mining and bayesian networks. These techniques are efficient in processing and mining information to identify claim behavior. Moreover, the practical capability was demonstrated in an Australian life insurance company, which supports business planning, development, and risk management (Chapter 4).

5. Natural language interactions (NLIs) enhanced by data visualization for customer behavior analysis from claim behavior is an emerging research direction. However, the design of visual analytics tools enhanced by NLIs for risk management and claim analysis requires thorough task analysis and domain expertise. Aiming to address such challenges, in this thesis, I investigate an alternative approach through a natural language interaction-based interactive visualization such as a chart, pie chart, or histogram, which can be used for insurance claim analysis and managing risk. This system supports analyzing multiple insurance datasets, such as relational, claim, and demographical. Experiment results show the performance of various input visualization components such as mouse, text, touch etc. (Chapter 5).

6. The practical applications of the proposed deep visual analytics solution is demonstrated to provide different outcome and visualize the classification performance in 3D projection with misclassified instances. The presented techniques and the discovered knowledge can benefit business stakeholders to understand customer behavior better and develop sustainable business industries (Chapter 6).

7.2 Future work

Although numerous data mining approaches and visual analytics (VA) have been applied to address crucial challenges in business problems, there is still more study and applications of data mining with visual analytics for customer behavior analysis to be done, as follows:

Chapter 3 introduced a pattern mining approach for discovering customer adverse behavior analysis, which considers the interaction between all possible adverse behavior patterns. I use a frequent pattern mining algorithm that can locate repeating relationships between unique items in a data set and represent them in association rules. Additionally, to compare with my proposed approach, I included a few outlier detection techniques such as Local Outlier Factor (LOF), Cluster-Based Local Outlier Factor (CBLOF), One-class SVM, and Isolation Forest (IF) to evaluate the performance of the proposed method. The experiment results on the life insurance data of 31,800 policyholders suggest that association rules can identify AS behavior and assist the insurance authority to reduce loss and guide changes to insurance premium policy for further development management and planning.

In contrast, there may be interactions between large criteria groups in real situations. It will be critical to determine if it is worth increasing the value of k, especially when there are many criteria under consideration. Therefore, in the future, since there are different outlier detection methods (e.g., clustering), I aim to combine these ways to develop a hybrid approach, e.g., using ensemble learning to integrate and understand the weights of different AS detection methods. I will also investigate the effects on the insurance industry after removing the months of lockdown and the change in behaviors and plans for the post-COVID-19 period.

Chapter 4 provided various visual interactive systems *ExVis, MHIVis* and *DiaVis* that have been carried out related to applying various evaluation techniques from different perspectives. However, there are still enough spaces that need to be discussed. Therefore, several potential future research directions are summarized: I will have a large dataset to display various factors for exploring information visualization. Also, it is possible to identify suspicious users, which domain experts would like to investigate. By integrating the advanced sequential pattern matching tool, the visual description of the framework can be expanded. Last, I would like to test the usefulness of *ExVis, MHIVis* and *DiaVis* with an eye-tracking model, which builds an understanding of human perception, cognition, and interaction to design environments.

The NLI-driven-DV method, presented in Chapter 5, is a general technique to explore customer claim behavior and manage risk. The analysis was performed solely based on the NLP approach on customer claim behavior data. It would be advantageous to add visual analytics to collaborative data analysis such as underwriting, mental health analysis, adverse selection, etc. While the N-Gram approach generates the visualization of multiple types of data and conveys how the corpus provides a better response, visual analytics requires advanced techniques such as transformer-based word embedding methods (BERT, RoBERTa etc.) and offer little variations in style. Thus, applying transformer-based word embedding methods may help address such limitations.

Chapter 6 provide the pattern mining and deep learning methods blending with visual interactive system analysis. In this chapter, various existing critical efforts have been carried out to apply various evaluation techniques from different perspectives. However, there are still enough spaces that need to be discussed. For example, it would be advantageous to improve the system to be used as a visual interactive DL tool to update and learn additional models at the instance level with a higher accuracy rate. I will also incorporate my system for text, binary, and visual data at a time.

In short, while this thesis provides a thorough report of my research in data mining techniques blending with visual analytics for customer behavior analysis, there are exciting and promising issues that remain unexplored. Therefore, I would like to continue studying and proposing practical techniques that bring insightful benefits to researchers in customer behavior analysis, especially for financial purposes. Additionally, I will also focus on immersive analytics (IA) in support of data visualization, a new research dimension that attempts to break through barriers between people, their data, and the analytical and decision-making tools they utilize. It has the potential to improve the visualization outcomes dramatically. For example, through IA, 3D technologies could provide new ways to use its position in space to the user and other data points. Although visualizing the results as an interactive 3D model improves the system's usability, IA virtual reality systems are one way for people to experience the materiality of the past by engaging with virtual representations of artifacts with 3D printed counterparts.

BIBLIOGRAPHY

[1] R. A. LEITE, T. GSCHWANDTNER, S. MIKSCH, E. GSTREIN, AND J. KUNTNER, *Neva: Visual analytics to identify fraudulent networks*, in Computer Graphics Forum, vol. 39, Wiley Online Library, 2020, pp. 344–359.

[2] M. ABADI, A. AGARWAL, P. BARHAM, E. BREVDO, Z. CHEN, C. CITRO, G. S. CORRADO, A. DAVIS, J. DEAN, M. DEVIN, ET AL., *Tensorflow: Large-scale machine learning on heterogeneous distributed systems*, arXiv preprint arXiv:1603.04467, (2016).

[3] A. AFROZ, L. ALI, M. N. KARIM, M. J. ALRAMADAN, K. ALAM, D. J. MAGLIANO, AND B. BILLAH, *Glycaemic control for people with type 2 diabetes mellitus in bangladesh-an urgent need for optimization of management plan*, Scientific reports, 9 (2019), pp. 1–10.

[4] S. AFZAL, S. GHANI, G. TISSINGTON, S. LANGODAN, H. P. DASARI, D. RAITSOS, J. GITTINGS, T. JAMIL, M. SRINIVASAN, AND I. HOTEIT, *Redseaatlas: A visual analytics tool for spatio-temporal multivariate data of the red sea.*, in EnvirVis@ EuroVis, 2019, pp. 25–32.

[5] E. AGICHTEIN AND Z. ZHENG, *Identifying" best bet" web search results by mining past user behavior*, in Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, 2006, pp. 902–908.

[6] R. AGRAWAL, T. IMIELIŃSKI, AND A. SWAMI, *Mining association rules between sets of items in large databases*, in Acm sigmod record, vol. 22, ACM, 1993, pp. 207–216.

[7] A. K. AHMAD, A. JAFAR, AND K. ALJOUMAA, *Customer churn prediction in telecom using machine learning in big data platform*, Journal of Big Data, 6 (2019), p. 28.

[8] F. AHMED, R. FERDOWS, M. R. ISLAM, AND A. R. M. KAMAL, *Deepvis: A visual interactive system for exploring performance of deep learning models*, in 2022 10th International Conference on Information and Education Technology (ICIET), IEEE, 2018.

[9] A. ALSAIARI, J. AURISANO, AND A. JOHNSON, *Evaluating strategies of exploratory visual data analysis in multi device environments*, (2020).

[10] A. AMIN, F. AL-OBEIDAT, B. SHAH, A. ADNAN, J. LOO, AND S. ANWAR, *Customer churn prediction in telecommunication industry using data certainty*, Journal of Business Research, 94 (2019), pp. 290–301.

[11] F. ANGIULLI AND C. PIZZUTI, *Outlier mining in large high-dimensional data sets*, IEEE transactions on Knowledge and Data engineering, 17 (2005), pp. 203–215.

[12] Z. ANWER, R. SHARMA, V. GARG, N. KUMAR, AND A. KUMARI, *Hypertension management in diabetic patients*, Eur Rev Med Pharmacol Sci, 15 (2011), pp. 1256–1263.

[13] A. APAOLAZA AND M. VIGO, *Assisted pattern mining for discovering interactive behaviours on the web*, International Journal of Human-Computer Studies, 130 (2019), pp. 196–208.

[14] K. AQUINO AND S. DOUGLAS, *Identity threat and antisocial behavior in organizations: The moderating effects of individual differences, aggressive modeling, and hierarchical status*, Organizational Behavior and Human Decision Processes, 90 (2003), pp. 195–208.

[15] C. ARAUZ-PACHECO, M. A. PARROTT, AND P. RASKIN, *The treatment of hypertension in adult patients with diabetes*, Diabetes care, 25 (2002), pp. 134–147.

[16] D. ARCHAMBAULT, H. PURCHASE, AND B. PINAUD, *Animation, small multiples, and the effect of mental map preservation in dynamic graphs*, IEEE transactions on visualization and computer graphics, 17 (2010), pp. 539–552.

[17] L. L. ARMSTRONG AND K. YOUNG, *Mind the gap: Person-centred delivery of mental health information to post-secondary students*, Psychosocial Intervention, 24 (2015), pp. 83–87.

[18] M. ARYCH AND W. DARCY, *General trends and competitiveness of australian life insurance industry*, Journal of International Studies, 13 (2020).

[19] M. Aupetit, M. Sedlmair, M. M. Abbas, A. Baggag, and H. Bensmail, *Toward perception-based evaluation of clustering techniques for visual analytics*, in 2019 IEEE Visualization Conference (VIS), IEEE, 2019, pp. 141–145.

[20] Australian Bureau of Statistics, *Socio-economic indexes for areas (SEIFA)*, 2016.

[21] A. Australian Bureau of Statistics, *National survey of mental health and wellbeing: Summary of results*, (Catalogue No. 4326.0), (2007).

[22] P. Bajari, C. Dalton, H. Hong, and A. Khwaja, *Moral hazard, adverse selection, and health expenditures: A semiparametric analysis*, The RAND Journal of Economics, 45 (2014), pp. 747–763.

[23] G. L. Bakris, M. Williams, L. Dworkin, W. J. Elliott, M. Epstein, R. Toto, K. Tuttle, J. Douglas, W. Hsueh, and J. Sowers, *Preserving renal function in adults with hypertension and diabetes: a consensus approach*, American journal of kidney diseases, 36 (2000), pp. 646–661.

[24] I. R. Baquedano, M. A. d. Santos, T. A. Martins, and M. L. Zanetti, *Self-care of patients with diabetes mellitus cared for at an emergency service in mexico*, Revista latino-americana de enfermagem, 18 (2010), pp. 1195–1202.

[25] R. Barcellos, J. Viterbo, F. Bernardini, and D. Trevisan, *An instrument for evaluating the quality of data visualizations*, in 2018 22nd International Conference Information Visualisation (IV), IEEE, 2018, pp. 169–174.

[26] D. W. Bates, D. J. Cullen, N. Laird, L. A. Petersen, S. D. Small, D. Servi, G. Laffel, B. J. Sweitzer, B. F. Shea, R. Hallisey, et al., *Incidence of adverse drug events and potential adverse drug events: implications for prevention*, Jama, 274 (1995), pp. 29–34.

[27] Z. Beasley, A. Friedman, L. Pieg, and P. Rosen, *Leveraging peer feedback to improve visualization education*, in 2020 IEEE Pacific Visualization Symposium (PacificVis), IEEE, 2020, pp. 146–155.

[28] C. Bharath, N. Saravanan, and S. Venkatalakshmi, *Assessment of knowledge related to diabetes mellitus among patients attending a dental college in salem city-a cross sectional study*, Brazilian Dental Science, 20 (2017), pp. 93–100.

[29] P. BHARDWAJ AND N. BALIYAN, *Hadoop based analysis and visualization of diabetes data through tableau*, in 2019 Twelfth International Conference on Contemporary Computing (IC3), IEEE, 2019, pp. 1–5.

[30] R. BIDDLE, S. LIU, P. TILOCCA, AND G. XU, *Automated underwriting in life insurance: Predictions and optimisation*, in Australasian Database Conference, Springer, 2018, pp. 135–146.

[31] R. BIDDLE, S. LIU, AND G. XU, *Semi-supervised soft k-means clustering of life insurance questionnaire responses*, in 2018 5th International Conference on Behavioral, Economic, and Socio-Cultural Computing (BESC), IEEE, 2018, pp. 30–31.

[32] T. BLASCHECK, K. KURZHALS, M. RASCHKE, M. BURCH, D. WEISKOPF, AND T. ERTL, *State-of-the-art of visualization for eye tracking data.*, in EuroVis (STARs), 2014.

[33] J. BOLHAAR, M. LINDEBOOM, AND B. VAN DER KLAAUW, *A dynamic analysis of the demand for health insurance and health care*, European Economic Review, 56 (2012), pp. 669–690.

[34] N. BOODHUN AND M. JAYABALAN, *Risk prediction in life insurance industry using supervised learning algorithms*, Complex & Intelligent Systems, 4 (2018), pp. 145–154.

[35] R. BORGO, L. MICALLEF, B. BACH, F. MCGEE, AND B. LEE, *Information visualization evaluation using crowdsourcing*, in Computer Graphics Forum, vol. 37, Wiley Online Library, 2018, pp. 573–595.

[36] R. BOURQUI, R. GIOT, AND D. AUBER, *Toward automatic comparison of visualization techniques: Application to graph visualization*, arXiv, (2019), pp. arXiv–1910.

[37] A. A. BOXWALA, J. KIM, J. M. GRILLO, AND L. OHNO-MACHADO, *Using statistical and machine learning to help institutions detect suspicious access to electronic health records*, Journal of the American Medical Informatics Association, 18 (2011), pp. 498–505.

[38] J. BUTLER, *Adverse selection in australian private health insurance*, in ACERH Policy Forum, 2007.

[39] J. J. CABAN AND D. GOTZ, *Visual analytics in healthcare–opportunities and research challenges*, 2015.

[40] L. CAO, *Behavior informatics and analytics: Let behavior talk*, in 2008 IEEE International Conference on Data Mining Workshops, IEEE, 2008, pp. 87–96.

[41] D. CASHMAN, G. PATTERSON, A. MOSCA, N. WATTS, S. ROBINSON, AND R. CHANG, *Rnnbow: Visualizing learning via backpropagation gradients in rnns*, IEEE Computer Graphics and Applications, 38 (2018), pp. 39–50.

[42] D. CENEDA, T. GSCHWANDTNER, T. MAY, S. MIKSCH, H.-J. SCHULZ, M. STREIT, AND C. TOMINSKI, *Characterizing guidance in visual analytics*, IEEE Transactions on Visualization and Computer Graphics, 23 (2016), pp. 111–120.

[43] J. CHAE, S. GAO, A. RAMANATHAN, C. A. STEED, AND G. TOURASSI, *Visualization for classification in deep neural networks*, tech. rep., Oak Ridge National Lab.(ORNL), Oak Ridge, TN (United States), 2017.

[44] C. CHANG, T. DWYER, AND K. MARRIOTT, *An evaluation of perceptually complementary views for multivariate data*, in 2018 IEEE Pacific Visualization Symposium (PacificVis), IEEE, 2018, pp. 195–204.

[45] R. CHANG, M. GHONIEM, R. KOSARA, W. RIBARSKY, J. YANG, E. SUMA, C. ZIEMKIEWICZ, D. KERN, AND A. SUDJIANTO, *Wirevis: Visualization of categorical, time-varying data from financial transactions*, in 2007 IEEE Symposium on Visual Analytics Science and Technology, IEEE, 2007, pp. 155–162.

[46] Y.-C. CHANG, C.-H. KU, AND C.-H. CHEN, *Using deep learning and visual analytics to explore hotel reviews and responses*, Tourism Management, 80 (2020), p. 104129.

[47] P. Y. CHAU, S. Y. HO, K. K. HO, AND Y. YAO, *Examining the effects of malfunctioning personalized services on online users' distrust and behaviors*, Decision Support Systems, 56 (2013), pp. 180–191.

[48] C.-Y. CHEN, *Using an eye tracker to investigate the effect of sticker on line app for older adults*, in International Conference on Human-Computer Interaction, Springer, 2019, pp. 225–234.

[49] M.-C. CHEN, A.-L. CHIU, AND H.-H. CHANG, *Mining changes in customer behavior in retail marketing*, Expert Systems with Applications, 28 (2005), pp. 773–781.

[50] N. CHEN, B. RIBEIRO, A. VIEIRA, AND A. CHEN, *Clustering and visualization of bankruptcy trajectory using self-organizing map*, Expert Systems with Applications, 40 (2013), pp. 385–393.

[51] Y. CHEN, Z. ZHENG, S. CHEN, L. SUN, AND D. CHEN, *Mining customer preference in physical stores from interaction behavior*, IEEE Access, 5 (2017), pp. 17436–17449.

[52] F. CHENG, Y. MING, AND H. QU, *Dece: Decision explorer with counterfactual explanations for machine learning models*, IEEE Transactions on Visualization and Computer Graphics, 27 (2020), pp. 1438–1447.

[53] F. CHENTLI, S. AZZOUG, AND S. MAHGOUN, *Diabetes mellitus in elderly*, Indian journal of endocrinology and metabolism, 19 (2015), p. 744.

[54] E. K. CHOE, B. LEE, H. ZHU, N. H. RICHE, AND D. BAUR, *Understanding self-reflection: how people reflect on personal data through visual data exploration*, in Proceedings of the 11th EAI International Conference on Pervasive Computing Technologies for Healthcare, 2017, pp. 173–182.

[55] J. CHOO AND S. LIU, *Visual analytics for explainable deep learning*, IEEE computer graphics and applications, 38 (2018), pp. 84–92.

[56] S.-F. CHOU, J.-S. HORNG, C.-H. S. LIU, AND J.-Y. LIN, *Identifying the critical factors of customer behavior: An integration perspective of marketing strategy and components of attitudes*, Journal of Retailing and Consumer Services, 55 (2020), p. 102113.

[57] I. CHOWDHURY, A. MOEID, E. HOQUE, M. A. KABIR, M. S. HOSSAIN, AND M. M. ISLAM, *MIVA: Multimodal interactions for facilitating visual analysis with multiple coordinated views*, in Proceedings of the 24th International Conference Information Visualisation, 2020, pp. 674–677.

[58] Y. A. CHRISTOBEL AND P. SIVAPRAKASAM, *A new classwise k nearest neighbor (cknn) method for the classification of diabetes dataset*, International Journal of Engineering and Advanced Technology, 2 (2013), pp. 396–200.

[59] A. M. CHU AND P. Y. CHAU, *Development and validation of instruments of information security deviant behavior*, Decision Support Systems, 66 (2014), pp. 93–101.

[60] S. CHUNG, S. SUH, C. PARK, K. KANG, J. CHOO, AND B. C. KWON, *Revacnn: Real-time visual analytics for convolutional neural network*, in KDD 16 Workshop on Interactive Data Exploration and Analytics, 2016.

[61] A. COHEN AND P. SIEGELMAN, *Testing for adverse selection in insurance markets*, Journal of Risk and insurance, 77 (2010), pp. 39–84.

[62] N. R. F. COLLABORATION ET AL., *Trends in adult body-mass index in 200 countries from 1975 to 2014: a pooled analysis of 1698 population-based measurement studies with 19· 2 million participants*, The lancet, 387 (2016), pp. 1377–1396.

[63] C. CONATI, G. CARENINI, E. HOQUE, B. STEICHEN, AND D. TOKER, *Evaluating the impact of user characteristics and different layouts on an interactive visualization for decision making*, in Computer Graphics Forum, vol. 33, Wiley Online Library, 2014, pp. 371–380.

[64] L. CONNELLY, *Logistic regression*, Medsurg Nursing, 29 (2020), pp. 353–354.

[65] L. COOK, *Mental health in australia: a quick guide*, Parliament of Australia, 14 (2019).

[66] A. COOPER AND K. V. PETRIDES, *A psychometric analysis of the trait emotional intelligence questionnaire–short form (teique–sf) using item response theory*, Journal of personality assessment, 92 (2010), pp. 449–457.

[67] K. COUSSEMENT AND K. W. DE BOCK, *Customer churn prediction in the online gambling industry: The beneficial effect of ensemble learning*, Journal of Business Research, 66 (2013), pp. 1629–1636.

[68] J. W. CRESWELL AND J. D. CRESWELL, *Research design: Qualitative, quantitative, and mixed methods approaches*, Sage publications, 2017.

[69] D. M. CUTLER AND R. J. ZECKHAUSER, *Adverse selection in health insurance*, in Forum for Health Economics & Policy, vol. 1, De Gruyter, 1998.

[70] W. Dai and M. G. Genton, *Multivariate functional data visualization and outlier detection*, Journal of Computational and Graphical Statistics, 27 (2018), pp. 923–934.

[71] A. Dasgupta, R. Kosara, and L. Gosink, *Vimtex: A visualization interface for multivariate, time-varying, geological data exploration*, in Computer Graphics Forum, vol. 34, Wiley Online Library, 2015, pp. 341–350.

[72] H. de Vries, A. Fishta, B. Weikert, A. R. Sanchez, and U. Wegewitz, *Determinants of sickness absence and return to work among employees with common mental disorders: a scoping review*, Journal of occupational rehabilitation, 28 (2018), pp. 393–417.

[73] F. Demir and B. Taşci, *An effective and robust approach based on r-cnn+ lstm model and ncar feature selection for ophthalmological disease detection from fundus images*, Journal of Personalized Medicine, 11 (2021), p. 1276.

[74] Z. Deng, D. Weng, J. Chen, R. Liu, Z. Wang, J. Bao, Y. Zheng, and Y. Wu, *Airvis: Visual analytics of air pollution propagation*, IEEE transactions on visualization and computer graphics, 26 (2019), pp. 800–810.

[75] S. K. Dey, A. Hossain, and M. M. Rahman, *Implementation of a web application to predict diabetes disease: an approach using machine learning algorithm*, in 2018 21st international conference of computer and information technology (ICCIT), IEEE, 2018, pp. 1–5.

[76] V. Dibia and Ç. Demiralp, *Data2vis: Automatic generation of data visualizations using sequence-to-sequence recurrent neural networks*, IEEE computer graphics and applications, 39 (2019), pp. 33–46.

[77] W. Didimo, L. Grilli, G. Liotta, L. Menconi, F. Montecchiani, and D. Pagliuca, *Combining network visualization and data mining for tax risk assessment*, IEEE Access, 8 (2020), pp. 16073–16086.

[78] Y. Ding, Y. Liu, H. Luan, and M. Sun, *Visualizing and understanding neural machine translation*, in Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2017, pp. 1150–1159.

[79] S. DONG, P. WANG, AND K. ABBAS, *A survey on deep learning and its applications*, Computer Science Review, 40 (2021), p. 100379.

[80] F. DU, C. PLAISANT, N. SPRING, AND B. SHNEIDERMAN, *Finding similar people to guide life choices: Challenge, design, and evaluation*, in Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, 2017, pp. 5498–5544.

[81] Y. DU, C. WANG, C. LI, AND H. YIN, *Behaviortracker: Visual analytics of customer switching behavior in o2o market*, in Proceedings of the 11th International Symposium on Visual Information Communication and Interaction, 2018, pp. 17–24.

[82] D. DUNBÄCK AND L. MATTSSON, *Predicting risk exposure in the insurance sector: Application of statistical tools to enhance price optimization at trygg-hansa*, 2021.

[83] J.-M. EKOE, R. GOLDENBERG, AND P. KATZ, *Screening for diabetes in adults*, Canadian journal of diabetes, 42 (2018), pp. S16–S19.

[84] N. ELMQVIST, J. STASKO, AND P. TSIGAS, *Datameadow: a visual canvas for analysis of large-scale multivariate data*, Information visualization, 7 (2008), pp. 18–33.

[85] M. J. EPPLER AND M. AESCHIMANN, *A systematic framework for risk visualization in risk management and communication*, Risk Management, 11 (2009), pp. 67–89.

[86] S. L. ETTNER, *Adverse selection and the purchase of medigap insurance by the elderly*, Journal of health economics, 16 (1997), pp. 543–562.

[87] A. FERRARIO, A. NOLL, AND M. V. WUTHRICH, *Insights from inside neural networks*, Available at SSRN 3226852, (2020).

[88] A. FINKELSTEIN, *Minimum standards, insurance regulation and adverse selection: evidence from the medigap market*, Journal of Public Economics, 88 (2004), pp. 2515–2547.

[89] T. FUJIWARA, N. SAKAMOTO, J. NONAKA, K. YAMAMOTO, K.-L. MA, ET AL., *A visual analytics framework for reviewing multivariate time-series data with*

dimensionality reduction, IEEE Transactions on Visualization and Computer Graphics, (2020).

[90] T. GAO, M. DONTCHEVA, E. ADAR, Z. LIU, AND K. G. KARAHALIOS, *Datatone: Managing ambiguity in natural language interfaces for data visualization*, in Proceedings of the 28th Annual ACM Symposium on User Interface Software & Technology, 2015, pp. 489–500.

[91] R. GARCIA, A. C. TELEA, B. C. DA SILVA, J. TØRRESEN, AND J. L. D. COMBA, *A task-and-technique centered survey on visual analytics for deep learning model engineering*, Computers & Graphics, 77 (2018), pp. 30–49.

[92] H. S. GARCIA CABALLERO, M. A. WESTENBERG, B. GEBRE, AND J. J. VAN WIJK, *V-awake: A visual analytics approach for correcting sleep predictions from deep learning models*, in Computer Graphics Forum, vol. 38, Wiley Online Library, 2019, pp. 1–12.

[93] S. GHEISARI, S. SHARIFLOU, J. PHU, P. J. KENNEDY, A. AGAR, M. KALLONIATIS, AND S. M. GOLZAN, *A combined convolutional and recurrent neural network for enhanced glaucoma detection*, Scientific reports, 11 (2021), pp. 1–11.

[94] A. GREWAL, M. KAUR, AND J. H. PARK, *A unified framework for behaviour monitoring and abnormality detection for smart home*, Wireless Communications and Mobile Computing, 2019 (2019).

[95] D. GUPTA, S. KHARE, AND A. AGGARWAL, *A method to predict diagnostic codes for chronic diseases using machine learning techniques*, in 2016 International Conference on Computing, Communication and Automation (ICCCA), IEEE, 2016, pp. 281–287.

[96] G. K. HADDAD AND M. Z. ANBAJI, *Analysis of adverse selection and moral hazard in the health insurance market of iran*, The Geneva Papers on Risk and Insurance-Issues and Practice, 35 (2010), pp. 581–599.

[97] H. HALEEM, Y. WANG, A. PURI, S. WADHWA, AND H. QU, *Evaluating the readability of force directed graph layouts: A deep learning approach*, IEEE computer graphics and applications, 39 (2019), pp. 40–53.

[98] C. Halkiopoulos, E. Gkintoni, and H. Antonopoulou, *Behavioral data analysis in emotional intelligence of social network consumers*, British Journal of Marketing Studies (BJMS), 8 (2), (2020), pp. 26–34.

[99] Y. Han, A. Rozga, J. Stasko, and G. D. Abowd, *Visual exploration of common behaviors for developmental health*, Visual Analytics in Healthcare, (2013).

[100] D. He, *The life insurance market: adverse selection revisited*, Economics Department, Washington University in St. Louis Campus, (2008).

[101] M. F. Hilton, P. A. Scuffham, N. Vecchio, and H. A. Whiteford, *Using the interaction of mental health symptoms and treatment status to estimate lost employee productivity*, Australian & New Zealand Journal of Psychiatry, 44 (2010), pp. 151–161.

[102] A. Hinterreiter, P. Ruch, H. Stitz, M. Ennemoser, J. Bernard, H. Strobelt, and M. Streit, *Confusionflow: A model-agnostic visualization for temporal analysis of classifier confusion*, IEEE Transactions on Visualization and Computer Graphics, (2020).

[103] J. C. Hoffmann, S. Mittal, C. H. Hoffmann, A. Fadl, A. Baadh, D. S. Katz, and J. Flug, *Combating the health risks of sedentary behavior in the contemporary radiology reading room*, American Journal of Roentgenology, 206 (2016), pp. 1135–1140.

[104] H. Höglund, *Tax payment default prediction using genetic algorithm-based variable selection*, Expert Systems with Applications, 88 (2017), pp. 368–375.

[105] F. Hohman, M. Kahng, R. Pienta, and D. H. Chau, *Visual analytics in deep learning: An interrogative survey for the next frontiers*, IEEE transactions on visualization and computer graphics, 25 (2018), pp. 2674–2693.

[106] F. Hohman, H. Park, C. Robinson, and D. H. P. Chau, *S ummit: Scaling deep learning interpretability by visualizing activation and attribution summarizations*, IEEE transactions on visualization and computer graphics, 26 (2019), pp. 1096–1106.

[107] F. Hohman, K. Wongsuphasawat, M. B. Kery, and K. Patel, *Understanding and visualizing data iteration in machine learning*, in Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, 2020, pp. 1–13.

[108] E. Hoque, S. Joty, L. Marquez, and G. Carenini, *Cqavis: Visual text analytics for community question answering*, in Proceedings of the 22nd International Conference on Intelligent User Interfaces, 2017, pp. 161–172.

[109] E. Hoque, V. Setlur, M. Tory, and I. Dykeman, *Applying pragmatics principles for interaction with visual analytics*, IEEE transactions on visualization and computer graphics, 24 (2017), pp. 309–318.

[110] K. Hu, M. A. Bakker, S. Li, T. Kraska, and C. Hidalgo, *Vizml: A machine learning approach to visualization recommendation*, in Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, 2019, pp. 1–12.

[111] T.-I. Hu and A. Tracogna, *Multichannel customer journeys and their determinants: Evidence from motor insurance*, Journal of Retailing and Consumer Services, 54 (2020), p. 102022.

[112] M. L. Huang, J. Liang, and Q. V. Nguyen, *A visualization approach for frauds detection in financial market*, in 2009 13th International Conference Information Visualisation, IEEE, 2009, pp. 197–202.

[113] Y. Huang and S. Meng, *Automobile insurance classification ratemaking based on telematics driving data*, Decision Support Systems, 127 (2019), p. 113156.

[114] B. J. Hutagaol and T. Mauritsius, *Risk level prediction of life insurance applicant using machine learning*, International Journal, 9 (2020).

[115] A. Iovine, F. Narducci, and G. Semeraro, *Conversational recommender systems and natural language:: A study through the converse framework*, Decision Support Systems, 131 (2020), p. 113250.

[116] M. Islam, I. Razzak, X. Wang, P. Tilocca, and G. Xu, *Natural language interactions enhanced by data visualization to explore insurance claims and manage risk*, Annals of Operations Research, (2021).

[117] M. R. Islam, S. Akter, M. R. Ratan, A. R. M. Kamal, and G. Xu, *Deep visual analytics (dva): Applications, challenges and future directions*, Human-Centric Intelligent Systems, 1 (2021), pp. 3–17.

[118] M. R. Islam, M. A. Kabir, A. Ahmed, A. R. M. Kamal, H. Wang, and A. Ulhaq, *Depression detection from social network data using machine learning techniques*, Health information science and systems, 6 (2018), pp. 1–12.

[119] M. R. Islam, A. R. M. Kamal, N. Sultana, R. Islam, M. A. Moni, et al., *Detecting depression using k-nearest neighbors (knn) classification technique*, in 2018 International Conference on Computer, Communication, Chemical, Material and Electronic Engineering (IC4ME2), IEEE, 2018, pp. 1–4.

[120] M. R. Islam, S. Liu, R. Biddle, I. Razzak, X. Wang, P. Tilocca, and G. Xu, *Discovering dynamic adverse behavior of policyholders in the life insurance industry*, Technological Forecasting and Social Change, 163 (2021), p. 120486.

[121] M. R. Islam, S. Liu, I. Razzak, M. A. Kabir, X. Wang, and G. Xu, *Mhivis: Visual analytics for exploring mental illness of policyholder's in life insurance industry*, in 2020 7th International Conference on Behavioral, Economic, and Socio-Cultural Computing (BESC), IEEE, 2020.

[122] M. R. Islam, S. Liu, X. Wang, and G. Xu, *Deep learning for misinformation detection on online social networks: a survey and new perspectives*, Social Network Analysis and Mining, 10 (2020), pp. 1–20.

[123] M. R. Islam, S. J. Miah, A. R. M. Kamal, and O. Burmeister, *A design construct of developing approaches to measure mental health conditions*, Australasian Journal of Information Systems, 23 (2019).

[124] M. R. Islam, I. Razzak, X. Wang, P. Tilocca, and G. Xu, *Ucbvis: understanding customer behavior sequences with visual interactive system*, in 2021 International Joint Conference on Neural Networks (IJCNN), IEEE, 2021, pp. 1–8.

[125] M. R. Islam, N. Sultana, M. A. Moni, P. C. Sarkar, and B. Rahman, *A comprehensive survey of time series anomaly detection in online social network data*, International Journal of Computer Applications, 180 (2017), pp. 13–22.

[126] M. R. Islam, J. Zhang, M. H. Ashmafee, I. Razzak, J. Zhou, X. Wang, and G. Xu, *Exvis: Explainable visual decision support system for risk management*, in International Conference on Behavioural and Social Computing, 2021.

[127] M. T. Islam, M. R. Islam, S. Akter, and M. Kawser, *Designing dashboard for exploring tourist hotspots in bangladesh*, in The 23th International Conference on Computer and Information Technology (ICCIT-2020, IEEE, 2020.

[128] A. JABLENSKY, *Treatment gaps and knowledge gaps in mental health: Schizophrenia as a global challenge*, Improving Mental Health Care, 403 (2013).

[129] C. JU, F. BAO, C. XU, AND X. FU, *A novel method of interestingness measures for association rules mining based on profit*, Discrete Dynamics in Nature and Society, 2015 (2015).

[130] K. JUDD, *Discrimination and insurance in australia*, Around Australia, (2011), p. 24.

[131] M. KAHNG, P. Y. ANDREWS, A. KALRO, AND D. H. CHAU, *A cti v is: Visual exploration of industry-scale deep neural network models*, IEEE transactions on visualization and computer graphics, 24 (2017), pp. 88–97.

[132] D. KALAIVANI AND P. SUMATHI, *Factor based prediction model for customer behavior analysis*, International Journal of System Assurance Engineering and Management, 10 (2019), pp. 519–524.

[133] R. R. KALYANI, S. H. GOLDEN, AND W. T. CEFALU, *Diabetes and aging: unique considerations and goals of care*, Diabetes Care, 40 (2017), pp. 440–443.

[134] I. KANDASAMY, W. V. KANDASAMY, J. M. OBBINENI, AND F. SMARANDACHE, *Indeterminate likert scale: feedback based on neutrosophy, its distance measures and clustering algorithm*, Soft Computing, 24 (2020), pp. 7459–7468.

[135] S. KAUSHIK AND C. GANDHI, *Ensure hierarchal identity based data security in cloud environment*, International Journal of Cloud Applications and Computing (IJCAC), 9 (2019), pp. 21–36.

[136] M. KEANE AND O. STAVRUNOVA,*Adverse selection, moral hazard and the demand for medigap insurance*, Journal of Econometrics, 190 (2016), pp. 62–78.

[137] D. KEIM, G. ANDRIENKO, J.-D. FEKETE, C. GÖRG, J. KOHLHAMMER, AND G. MELANÇON, *Visual analytics: Definition, process, and challenges*, in Information visualization, Springer, 2008, pp. 154–175.

[138] A. A. KHADE, *Performing customer behavior analysis using big data analytics*, Procedia computer science, 79 (2016), pp. 986–992.

[139] A. M. KILBOURNE, K. BECK, B. SPAETH-RUBLEE, P. RAMANUJ, R. W. O'BRIEN, N. TOMOYASU, AND H. A. PINCUS, *Measuring and improving the quality of mental health care: a global perspective*, World psychiatry, 17 (2018), pp. 30–38.

[140] J. K. KIM, H. S. SONG, T. S. KIM, AND H. K. KIM, *Detecting the change of customer behavior based on decision tree analysis*, Expert Systems, 22 (2005), pp. 193–205.

[141] M. S. KIRKMAN, V. J. BRISCOE, N. CLARK, H. FLOREZ, L. B. HAAS, J. B. HALTER, E. S. HUANG, M. T. KORYTKOWSKI, M. N. MUNSHI, P. S. ODEGARD, ET AL., *Diabetes in older adults*, Diabetes care, 35 (2012), pp. 2650–2664.

[142] A. K. KNUDSEN, S. ØVERLAND, M. HOTOPF, AND A. MYKLETUN, *Lost working years due to mental disorders: an analysis of the norwegian disability pension registry*, (2012).

[143] S. KO, I. CHO, S. AFZAL, C. YAU, J. CHAE, A. MALIK, K. BECK, Y. JANG, W. RIBARSKY, AND D. S. EBERT, *A survey on visual analysis approaches for financial data*, in Computer Graphics Forum, vol. 35, Wiley Online Library, 2016, pp. 599–617.

[144] S. KOLDIJK, J. BERNARD, T. RUPPERT, J. KOHLHAMMER, M. NEERINCX, AND W. KRAAIJ, *Visual analytics of work behavior data-insights on individual differences*, (2015).

[145] I. V. KRAK, O. V. BARMAK, AND E. MANZIUK, *Visual analytics to build a machine learning model*, in Research Advancements in Smart Technology, Optimization, and Renewable Energy, IGI Global, 2021, pp. 313–329.

[146] A. KRIZHEVSKY, G. HINTON, ET AL., *Learning multiple layers of features from tiny images*, (2009).

[147] C. H. KU, Y.-C. CHANG, Y. WANG, C.-H. CHEN, AND S.-H. HSIAO, *Artificial intelligence and visual analytics: A deep-learning approach to analyze hotel reviews & responses*, in Proceedings of the 52nd Hawaii International Conference on System Sciences, 2019.

[148] T. KULESZA, M. BURNETT, W.-K. WONG, AND S. STUMPF, *Principles of explanatory debugging to personalize interactive machine learning*, in Proceedings of

the 20th international conference on intelligent user interfaces, 2015, pp. 126–137.

[149] B. C. KWON, V. ANAND, K. A. SEVERSON, S. GHOSH, Z. SUN, B. I. FROHNERT, M. LUNDGREN, AND K. NG, *Dpvis: Visual analytics with hidden markov models for disease progression pathways*, IEEE transactions on visualization and computer graphics, (2020).

[150] B. C. KWON, M.-J. CHOI, J. T. KIM, E. CHOI, Y. B. KIM, S. KWON, J. SUN, AND J. CHOO, *Retainvis: Visual analytics with interpretable and interactive recurrent neural networks on electronic medical records*, IEEE transactions on visualization and computer graphics, 25 (2018), pp. 299–309.

[151] B. C. KWON, J. VERMA, AND A. PERER, *Peekquence: Visual analytics for event sequence data*, in ACM SIGKDD 2016 Workshop on Interactive Data Exploration and Analytics, vol. 1, 2016.

[152] T. LANG AND M. RETTENMEIER, *Understanding consumer behavior with recurrent neural networks*, in Workshop on Machine Learning Methods for Recommender Systems, 2017.

[153] M. T. R. LASKAR, E. HOQUE, AND J. HUANG, *Query focused abstractive summarization via incorporating query relevance and transfer learning with transformer models*, in Canadian Conference on Artificial Intelligence, Springer, 2020, pp. 342–348.

[154] P.-M. LAW, Z. LIU, S. MALIK, AND R. C. BASOLE, *Maqui: Interweaving queries and pattern mining for recursive event sequence exploration*, IEEE transactions on visualization and computer graphics, 25 (2018), pp. 396–406.

[155] L. S. LEACH, P. BUTTERWORTH, AND H. WHITEFORD, *Private health insurance, mental health and service use in australia*, Australian & New Zealand Journal of Psychiatry, 46 (2012), pp. 468–475.

[156] B. J. LEE, B. KU, J. NAM, D. D. PHAM, AND J. Y. KIM, *Prediction of fasting plasma glucose status using anthropometric measures for diagnosing type 2 diabetes*, IEEE journal of biomedical and health informatics, 18 (2013), pp. 555–561.

[157] C. LEE, Y. KIM, S. JIN, D. KIM, R. MACIEJEWSKI, D. EBERT, AND S. KO, *A visual analytics system for exploring, monitoring, and forecasting road traffic*

congestion, IEEE Transactions on Visualization and Computer Graphics, 26 (2020), pp. 3133–3146.

[158] S. LEE, C. MIN, C. YOO, AND J. SONG, *Understanding customer malling behavior in an urban shopping mall using smartphones*, in Proceedings of the 2013 ACM conference on Pervasive and ubiquitous computing adjunct publication, 2013, pp. 901–910.

[159] R. A. LEITE, T. GSCHWANDTNER, S. MIKSCH, E. GSTREIN, AND J. KUNTNER, *Visual analytics for fraud detection and monitoring*, in 2015 IEEE Conference on Visual Analytics Science and Technology (VAST), IEEE, 2015, pp. 201–202.

[160] R. A. LEITE, T. GSCHWANDTNER, S. MIKSCH, S. KRIGLSTEIN, M. POHL, E. GSTREIN, AND J. KUNTNER, *Eva: Visual analytics to identify fraudulent events*, IEEE transactions on visualization and computer graphics, 24 (2017), pp. 330–339.

[161] B. LESTER, A. SHOURIDEH, V. VENKATESWARAN, AND A. ZETLIN-JONES, *Screening and adverse selection in frictional markets*, Journal of Political Economy, 127 (2019), pp. 338–377.

[162] D. LI, L. DENG, B. B. GUPTA, H. WANG, AND C. CHOI, *A novel cnn based security guaranteed image watermarking generation scenario for smart city applications*, Information Sciences, 479 (2019), pp. 432–447.

[163] K. LI, Y.-N. LI, H. YIN, Y. HU, P. YE, AND C. WANG, *Visual analysis of retailing store location selection*, Journal of Visualization, 23 (2020), pp. 1071–1086.

[164] Q. LI, P. SCHAFFER, J. PANG, AND S. MAUW, *Comparative analysis of clustering protocols with probabilistic model checking*, in 2012 Sixth International Symposium on Theoretical Aspects of Software Engineering, IEEE, 2012, pp. 249–252.

[165] Q. LI AND Z. WANG, *Riemannian submanifold tracking on low-rank algebraic variety*, in Thirty-First AAAI Conference on Artificial Intelligence, 2017.

[166] R. LI, C. YIN, S. YANG, B. QIAN, AND P. ZHANG, *Marrying medical domain knowledge with deep learning on electronic health records: A deep visual analytics approach*, Journal of Medical Internet Research, 22 (2020), p. e20645.

[167] C. LIN, Y.-J. HSIAO, AND C.-Y. YEH, *Financial literacy, financial advisors, and information sources on demand for life insurance*, Pacific-Basin Finance Journal, 43 (2017), pp. 218–237.

[168] C. LIN, C. M. LIN, B. LIN, AND M.-C. YANG, *A decision support system for improving doctors' prescribing behavior*, Expert Systems with Applications, 36 (2009), pp. 7975–7984.

[169] G. S. LINOFF AND M. J. BERRY, *Data mining techniques: for marketing, sales, and customer relationship management*, John Wiley & Sons, 2011.

[170] Z. C. LIPTON, *The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery.*, Queue, 16 (2018), pp. 31–57.

[171] H. LIU, T. TANIGUCHI, Y. TANAKA, K. TAKENAKA, AND T. BANDO, *Visualization of driving behavior based on hidden feature extraction by using deep learning*, IEEE Transactions on Intelligent Transportation Systems, 18 (2017), pp. 2477–2489.

[172] S. LIU, G. XU, X. ZHU, AND Z. ZHOU, *Towards simplified insurance application via sparse questionnaire optimization*, in 2017 International Conference on Behavioral, Economic, Socio-Cultural Computing (BESC), IEEE, 2017, pp. 1–2.

[173] Z. LIU, Y. WANG, M. DONTCHEVA, M. HOFFMAN, S. WALKER, AND A. WILSON, *Patterns and sequences: Interactive exploration of clickstreams to understand common visitor paths*, IEEE Transactions on Visualization and Computer Graphics, 23 (2016), pp. 321–330.

[174] V. E.-W. LO AND P. A. GREEN, *Development and evaluation of automotive speech interfaces: useful information from the human factors and the related literature*, International Journal of Vehicular Technology, 2013 (2013).

[175] Y. LU, R. GARCIA, B. HANSEN, M. GLEICHER, AND R. MACIEJEWSKI, *The state-of-the-art in predictive visual analytics*, in Computer Graphics Forum, vol. 36, Wiley Online Library, 2017, pp. 539–562.

[176] N. MAHAN, S. JHA, AND R. SWANSON, *Employing visual analytics to understand worldwide prevalence and impact of diabetes epidemic*, (2017).

[177] S. T. Mai, X. He, J. Feng, C. Plant, and C. Böhm, *Anytime density-based clustering of complex data*, Knowledge and Information Systems, 45 (2015), pp. 319–355.

[178] A. Malik, R. Maciejewski, N. Elmqvist, Y. Jang, D. S. Ebert, and W. Huang, *A correlative analysis process in a visual analytics environment*, in 2012 IEEE Conference on Visual Analytics Science and Technology (VAST), IEEE, 2012, pp. 33–42.

[179] A. Mandal, A. Sinaeepourfard, and S. K. Naskar, *Vda: Deep learning based visual data analysis in integrated edge to cloud computing environment*, in Adjunct Proceedings of the 2021 International Conference on Distributed Computing and Networking, 2021, pp. 7–12.

[180] M. Maniruzzaman, M. J. Rahman, B. Ahammed, and M. M. Abedin, *Classification and prediction of diabetes disease using machine learning paradigm*, Health information science and systems, 8 (2020), pp. 1–14.

[181] D. McCarthy and O. S. Mitchell, *International adverse selection in life insurance and annuities*, in Ageing in advanced industrial states, Springer, 2010, pp. 119–135.

[182] G. Meyer, G. Adomavicius, P. E. Johnson, M. Elidrisi, W. A. Rush, J. M. Sperl-Hillen, and P. J. O'Connor, *A machine learning approach to improving dynamic decision making*, Information Systems Research, 25 (2014), pp. 239–263.

[183] T. E. Meyer, M. Monroe, C. Plaisant, R. Lan, K. Wongsuphasawat, T. S. Coster, S. Gold, J. Millstein, and B. Shneiderman, *Visualizing patterns of drug prescriptions with eventflow: A pilot study of asthma medications in the military health system*, tech. rep., OFFICE OF THE SURGEON GENERAL (ARMY) FALLS CHURCH VA, 2013.

[184] J. Mez, D. H. Daneshvar, P. T. Kiernan, B. Abdolmohammadi, V. E. Alvarez, B. R. Huber, M. L. Alosco, T. M. Solomon, C. J. Nowinski, L. McHale, et al., *Clinicopathological evaluation of chronic traumatic encephalopathy in players of american football*, Jama, 318 (2017), pp. 360–370.

[185] Y. MING, S. CAO, R. ZHANG, Z. LI, Y. CHEN, Y. SONG, AND H. QU, *Understanding hidden memories of recurrent neural networks*, in 2017 IEEE Conference on Visual Analytics Science and Technology (VAST), IEEE, 2017, pp. 13–24.

[186] D. C. MOHR, M. N. BURNS, S. M. SCHUELLER, G. CLARKE, AND M. KLINKMAN, *Behavioral intervention technologies: evidence review and recommendations for future research in mental health*, General hospital psychiatry, 35 (2013), pp. 332–338.

[187] G. MONTAVON, W. SAMEK, AND K.-R. MÜLLER, *Methods for interpreting and understanding deep neural networks*, Digital Signal Processing, 73 (2018), pp. 1–15.

[188] A. A. MUBARAK, H. CAO, W. ZHANG, AND W. ZHANG, *Visual analytics of video-clickstream data and prediction of learners' performance using deep learning models in moocs' courses*, Computer Applications in Engineering Education, (2020).

[189] J. MÜLLER, M. CYPKO, A. OESER, M. STOEHR, V. ZEBRALLA, S. SCHREIBER, S. WIEGAND, A. DIETZ, AND S. OELTZE-JAFRA, *Visual assistance in clinical decision support*, (2021).

[190] J. MÜLLER, M. STOEHR, A. OESER, J. GAEBEL, M. STREIT, A. DIETZ, AND S. OELTZE-JAFRA, *A visual approach to explainable computerized clinical decision support*, Computers & Graphics, 91 (2020), pp. 1–11.

[191] P. MUNTNER, L. D. COLANTONIO, M. CUSHMAN, D. C. GOFF, G. HOWARD, V. J. HOWARD, B. KISSELA, E. B. LEVITAN, D. M. LLOYD-JONES, AND M. M. SAFFORD, *Validation of the atherosclerotic cardiovascular disease pooled cohort risk equations*, Jama, 311 (2014), pp. 1406–1415.

[192] J. NAHAR, T. IMAM, K. S. TICKLE, AND Y.-P. P. CHEN, *Association rule mining to detect factors which contribute to heart disease in males and females*, Expert Systems with Applications, 40 (2013), pp. 1086–1093.

[193] A. NARECHANIA, A. SRINIVASAN, AND J. STASKO, *Nl4dv: A toolkit for generating analytic specifications for data visualization from natural language queries*, arXiv preprint arXiv:2008.10723, (2020).

[194] E. W. NGAI, L. XIU, AND D. C. CHAU, *Application of data mining techniques in customer relationship management: A literature review and classification*, Expert systems with applications, 36 (2009), pp. 2592–2602.

[195] A. NGUYEN, J. YOSINSKI, AND J. CLUNE, *Multifaceted feature visualization: Uncovering the different types of features learned by each neuron in deep neural networks*, arXiv preprint arXiv:1602.03616, (2016).

[196] L. NGUYEN AND A. WORTHINGTON, *Adverse selection in private health insurance*, Consumer Interests Annual, 63 (2017), pp. 25–34.

[197] Z. NIU, D. CHENG, L. ZHANG, AND J. ZHANG, *Visual analytics for networked-guarantee loans risk management*, in 2018 IEEE Pacific Visualization Symposium (PacificVis), IEEE, 2018, pp. 160–169.

[198] J. OBEID AND E. HOQUE, *Chart-to-text: Generating natural language descriptions for charts by adapting the transformer model*, arXiv preprint arXiv:2010.09142, (2020).

[199] O. O. OLAKANMI AND A. DADA, *An efficient privacy-preserving approach for secure verifiable outsourced computing on untrusted platforms*, International Journal of Cloud Applications and Computing (IJCAC), 9 (2019), pp. 79–98.

[200] L. ORDONEZ-ANTE, G. VAN SEGHBROECK, T. WAUTERS, B. VOLCKAERT, AND F. DE TURCK, *Explora: Interactive querying of multidimensional data in the context of smart cities*, Sensors, 20 (2020), p. 2737.

[201] M. J. PAGE, J. E. MCKENZIE, P. M. BOSSUYT, I. BOUTRON, T. C. HOFFMANN, C. D. MULROW, L. SHAMSEER, J. M. TETZLAFF, E. A. AKL, S. E. BRENNAN, ET AL., *The prisma 2020 statement: an updated guideline for reporting systematic reviews*, International Journal of Surgery, 88 (2021), p. 105906.

[202] C. PARK, H. KIM, AND K. LEE, *A visualization system for performance analysis of image classification models*, Electronic Imaging, 2020 (2020), pp. 375–1.

[203] M. V. PAULY AND Y. ZENG, *Adverse selection and the challenges to stand-alone prescription drug insurance*, in Forum for Health Economics & Policy, vol. 7, De Gruyter, 2004.

[204] A. Z. Peixinho, B. C. Benato, L. G. Nonato, and A. X. Falcão, *Delaunay triangulation data augmentation guided by visual analytics for deep learning*, in 2018 31st SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI), IEEE, 2018, pp. 384–391.

[205] J. Pesantez-Narvaez, M. Guillen, and M. Alcañiz, *Predicting motor insurance claims using telematics data xgboost versus logistic regression*, Risks, 7 (2019), p. 70.

[206] N. Pezzotti, T. Höllt, J. Van Gemert, B. P. Lelieveldt, E. Eisemann, and A. Vilanova, *Deepeyes: Progressive visual analytics for designing deep neural networks*, IEEE transactions on visualization and computer graphics, 24 (2017), pp. 98–108.

[207] B. Pink, *Socio-economic indexes for areas (seifa)*, Canberra: Australian Bureau of Statistics, (2011).

[208] M. Polyakova, *Regulation of insurance with adverse selection and switching costs: Evidence from medicare part d*, American Economic Journal: Applied Economics, 8 (2016), pp. 165–95.

[209] R. Puelz and A. Snow, *Evidence on adverse selection: Equilibrium signaling and cross-subsidization in the insurance market*, Journal of Political Economy, 102 (1994), pp. 236–257.

[210] X. Qian, R. A. Rossi, F. Du, S. Kim, E. Koh, S. Malik, T. Y. Lee, and J. Chan, *Ml-based visualization recommendation: Learning to recommend visualizations from data*, arXiv preprint arXiv:2009.12316, (2020).

[211] ——, *Learning to recommend visualizations from data*, in Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, 2021, pp. 1359–1369.

[212] Z. Qu, C. W. Lau, D. R. Catchpoole, S. Simoff, and Q. V. Nguyen, *Intelligent and immersive visual analytics of health data*, in Advanced Computational Intelligence in Healthcare-7, Springer, 2020, pp. 29–44.

[213] M. Rahman, M. R. Islam, S. Akter, S. Akter, L. Islam, and G. Xu, *Diavis: Exploration and analysis of diabetes through visual interactive system*, Human-Centric Intelligent Systems.

[214] I. Razzak, R. A. Saris, M. Blumenstein, and G. Xu, *Integrating joint feature selection into subspace learning: A formulation of 2dpca for outliers robust feature selection*, Neural Networks, 121 (2020), pp. 441–451.

[215] I. Razzak, K. Zafar, M. Imran, and G. Xu, *Randomized nonlinear one-class support vector machines with bounded loss function to detect of outliers for large scale iot data*, Future Generation Computer Systems, 112 (2020), pp. 715–723.

[216] M. Riddel and D. Hales, *Risk misperceptions and selection in insurance markets: An application to demand for cancer insurance*, Journal of Risk and Insurance, 85 (2018), pp. 749–785.

[217] A. Rind, P. Federico, T. Gschwandtner, W. Aigner, J. Doppler, and M. Wagner, *Visual analytics of electronic health records with a focus on time*, in New Perspectives in Medical Records, Springer, 2017, pp. 65–77.

[218] J. C. Roberts, *State of the art: Coordinated & multiple views in exploratory visualization*, in Fifth international conference on coordinated and multiple views in exploratory visualization (CMV 2007), IEEE, 2007, pp. 61–71.

[219] C. Roelen, G. Norder, P. Koopmans, W. Van Rhenen, J. Van Der Klink, and U. Bültmann, *Employees sick-listed with mental disorders: who returns to work and when?*, Journal of occupational rehabilitation, 22 (2012), pp. 409–417.

[220] M. Röhlig, O. Stachs, and H. Schumann, *Detection of diabetic neuropathy-can visual analytics methods really help in practice?*, in EuroRV^3@ EuroVis, 2016, pp. 19–21.

[221] X. Rong and E. Adar, *Visual tools for debugging neural language models*, in Proceedings of ICML Workshop on Visualization for Deep Learning, 2016.

[222] S. Rosenbaum, *Insurance discrimination on the basis of health status: An overview of discrimination practices, federal law, and federal reform options: Executive summary*, Journal of Law, Medicine & Ethics, 37 (2009), pp. 101–120.

[223] S. K. Roy, R. L. Gruner, and J. Guo, *Exploring customer experience, commitment, and engagement behaviours*, Journal of Strategic Marketing, (2020), pp. 1–24.

[224] S. RUDOLPH, A. SAVIKHIN, AND D. S. EBERT, *Finvis: Applied visual analytics for personal financial planning*, in 2009 IEEE symposium on visual analytics science and technology, IEEE, 2009, pp. 195–202.

[225] L. M. RUILOPE AND R. GARCÍA-ROBLES, *How far should blood pressure be reduced in diabetic hypertensive patients?*, Journal of Hypertension, 15 (1997), pp. S63–S65.

[226] D. SACHA, M. KRAUS, D. A. KEIM, AND M. CHEN, *Vis4ml: An ontology for visual analytics assisted machine learning*, IEEE transactions on visualization and computer graphics, 25 (2018), pp. 385–395.

[227] D. SACHA, M. SEDLMAIR, L. ZHANG, J. A. LEE, D. WEISKOPF, S. NORTH, AND D. KEIM, *Human-centered machine learning through interactive visualization*, ESANN, 2016.

[228] K. SAHU, Y. BAI, AND Y. CHOI, *Supervised sentiment analysis of twitter handle of president trump with data visualization technique*, in 2020 10th Annual Computing and Communication Workshop and Conference (CCWC), IEEE, 2020, pp. 0640–0646.

[229] Z. SAJID, F. KHAN, AND Y. ZHANG, *Integration of interpretive structural modelling with bayesian network for biodiesel performance analysis*, Renewable Energy, 107 (2017), pp. 194–203.

[230] W. SAMEK, A. BINDER, G. MONTAVON, S. LAPUSCHKIN, AND K.-R. MÜLLER, *Evaluating the visualization of what a deep neural network has learned*, IEEE transactions on neural networks and learning systems, 28 (2016), pp. 2660–2673.

[231] W. SAMEK, T. WIEGAND, AND K.-R. MÜLLER, *Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models*, arXiv preprint arXiv:1708.08296, (2017).

[232] J. SAMUEL, G. ALI, M. RAHMAN, E. ESAWI, Y. SAMUEL, ET AL., *Covid-19 public sentiment insights and machine learning for tweets classification*, Information, 11 (2020), p. 314.

[233] N. Saquib, M. A. Khanam, J. Saquib, S. Anand, G. M. Chertow, M. Barry, T. Ahmed, and M. R. Cullen, *High prevalence of type 2 diabetes among the urban middle class in bangladesh*, BMC public health, 13 (2013), pp. 1–9.

[234] M. A. Sarwar, N. Kamal, W. Hamid, and M. A. Shah, *Prediction of diabetes using machine learning algorithms in healthcare*, in 2018 24th International Conference on Automation and Computing (ICAC), IEEE, 2018, pp. 1–6.

[235] A. Satyanarayan, D. Moritz, K. Wongsuphasawat, and J. Heer, *Vega-lite: A grammar of interactive graphics*, IEEE transactions on visualization and computer graphics, 23 (2016), pp. 341–350.

[236] J. Schelldorfer and M. V. Wuthrich, *Nesting classical actuarial models into neural networks*, Available at SSRN 3320525, (2019).

[237] H.-J. Schulz, M. Angelini, G. Santucci, and H. Schumann, *An enhanced visualization process model for incremental visualization*, IEEE transactions on visualization and computer graphics, 22 (2015), pp. 1830–1842.

[238] C. Seifert, A. Aamir, A. Balagopalan, D. Jain, A. Sharma, S. Grottel, and S. Gumhold, *Visualizations of deep neural networks in computer vision: A survey*, in Transparent data mining for big and small data, Springer, 2017, pp. 123–144.

[239] R. Sengupta and D. Rooj, *The effect of health insurance on hospitalization: Identification of adverse selection, moral hazard and the vulnerable population in the indian healthcare market*, World Development, 122 (2019), pp. 110–129.

[240] V. Setlur, S. E. Battersby, M. Tory, R. Gossweiler, and A. X. Chang, *Eviza: A natural language interface for visual analysis*, in Proceedings of the 29th Annual Symposium on User Interface Software and Technology, 2016, pp. 365–377.

[241] L. Shao, N. Silva, E. Eggeling, and T. Schreck, *Visual exploration of large scatter plot matrices by pattern recommendation based on eye tracking*, in Proceedings of the 2017 ACM Workshop on Exploratory Search and Interactive Data Analytics, 2017, pp. 9–16.

[242] O. Sharif, M. R. Islam, M. Z. Hasan, M. A. Kabir, M. E. Hasan, S. A. AlQahtani, and G. Xu, *Analyzing the impact of demographic variables on spreading*

and forecasting covid-19, Journal of Healthcare Informatics Research, (2021), pp. 1–19.

[243] D. SHIN, S. HE, G. M. LEE, A. B. WHINSTON, S. CETINTAS, AND K.-C. LEE, *Enhancing social media analysis with visual data analytics: A deep learning approach.*, MIS Quarterly, 44 (2020).

[244] K. SINGH AND P. BEST, *Anti-money laundering: Using data visualization to identify suspicious activity*, International Journal of Accounting Information Systems, 34 (2019), p. 100418.

[245] N. SINGH AND M. VARDHAN, *Distributed ledger technology based property transaction system with support for iot devices*, International Journal of Cloud Applications and Computing (IJCAC), 9 (2019), pp. 60–78.

[246] X.-P. SONG, Z.-H. HU, J.-G. DU, AND Z.-H. SHENG, *Application of machine learning methods to risk assessment of financial statement fraud: evidence from china*, Journal of Forecasting, 33 (2014), pp. 611–626.

[247] A. SORIANO-VARGAS, B. HAMANN, AND M. C. F DE OLIVEIRA, *Tv-mv analytics: A visual analytics framework to explore time-varying multivariate data*, Information Visualization, 19 (2020), pp. 3–23.

[248] J. L. SPEARS AND H. BARKI, *User participation in information systems security risk management*, MIS quarterly, (2010), pp. 503–522.

[249] A. SRINIVASAN, B. LEE, N. HENRY RICHE, S. M. DRUCKER, AND K. HINCKLEY, *Inchorus: Designing consistent multimodal interactions for data visualization on tablet devices*, in Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, 2020, pp. 1–13.

[250] A. SRINIVASAN AND J. STASKO, *Natural language interfaces for data analysis with visualization: Considering what has and could be asked*, in Proceedings of the Eurographics/IEEE VGTC Conference on Visualization: Short Papers, 2017, pp. 55–59.

[251] ——, *How to ask what to say?: Strategies for evaluating natural language interfaces for data visualization*, IEEE Computer Graphics and Applications, 40 (2020), pp. 96–103.

[252] S. STEHLE AND R. KITCHIN, *Real-time and archival data visualisation techniques in city dashboards*, International Journal of Geographical Information Science, 34 (2020), pp. 344–366.

[253] C. STEYN, C. DAVIES, AND A. SAMBO, *Eliciting student feedback for course development: the application of a qualitative course evaluation tool among business research students*, Assessment & Evaluation in Higher Education, 44 (2019), pp. 11–24.

[254] C. STOLTE, D. TANG, AND P. HANRAHAN, *Polaris: A system for query, analysis, and visualization of multidimensional relational databases*, IEEE Transactions on Visualization and Computer Graphics, 8 (2002), pp. 52–65.

[255] H. STROBELT, S. GEHRMANN, H. PFISTER, AND A. M. RUSH, *Lstmvis: A tool for visual analysis of hidden state dynamics in recurrent neural networks*, IEEE transactions on visualization and computer graphics, 24 (2017), pp. 667–676.

[256] T. SU, Z. CAO, Z. LV, C. LIU, AND X. LI, *Multi-dimensional visualization of large-scale marine hydrological environmental data*, Advances in Engineering Software, 95 (2016), pp. 7–15.

[257] Y. SUN, J. LEIGH, A. JOHNSON, AND S. LEE, *Articulate: A semi-automated model for translating natural language queries into meaningful visualizations*, in International Symposium on Smart Graphics, Springer, 2010, pp. 184–195.

[258] S. SWAID, M. MAAT, H. KRISHNAN, D. GHOSHAL, AND L. RAMAKRISHNAN, *Usability heuristic evaluation of scientific data analysis and visualization tools*, in International Conference on Applied Human Factors and Ergonomics, Springer, 2017, pp. 471–482.

[259] V. SWAMINATHAN AND R. SIVAKUMAR, *A comparative study of recent ontology visualization tools with a case of diabetes data*, International Journal of Research in Computer Science, 2 (2012), p. 31.

[260] F. TAGHIKHAH, A. VOINOV, N. SHUKLA, AND T. FILATOVA, *Exploring consumer behavior and policy options in organic food adoption: Insights from the australian wine sector*, Environmental Science & Policy, 109 (2020), pp. 116–124.

[261] K. TAKAI AND K. YADA, *A framework for analysis of the effect of time on shopping behavior*, Journal of Intelligent Information Systems, 41 (2013), pp. 91–107.

[262] T. TAKENAKA, *Analysis and prediction of customer behaviors for restaurant management*, in Service Engineering for Gastronomic Sciences, Springer, 2020, pp. 29–41.

[263] A. TEWARI AND B. GUPTA, *Security, privacy and trust of different layers in internet-of-things (iots) framework*, Future generation computer systems, 108 (2020), pp. 909–920.

[264] D. P. THOMAS, R. BORGO, R. S. LARAMEE, AND S. HANDS, *Qcdvis: a tool for the visualisation of quantum chromodynamics (qcd) data*, in Proceedings of the 33rd Spring Conference on Computer Graphics, 2017, pp. 1–14.

[265] J. THOMPSON, A. SRINIVASAN, AND J. STASKO, *Tangraphe: interactive exploration of network visualizations using single hand, multi-touch gestures*, in Proceedings of the 2018 International Conference on Advanced Visual Interfaces, 2018, pp. 1–5.

[266] F. TIAN, T. LAN, K.-M. CHAO, N. GODWIN, Q. ZHENG, N. SHAH, AND F. ZHANG, *Mining suspicious tax evasion groups in big data*, IEEE Transactions on Knowledge and Data Engineering, 28 (2016), pp. 2651–2664.

[267] Z. TIAN, Y. ZHENG, X. LI, S. DU, AND X. XU, *Graph convolutional network based optic disc and cup segmentation on fundus images*, Biomedical Optics Express, 11 (2020), pp. 3043–3057.

[268] J. TOSADO, L. ZDILAR, H. ELHALAWANI, B. ELGOHARI, D. M. VOCK, G. E. MARAI, C. FULLER, A. S. MOHAMED, AND G. CANAHUATE, *clustering of largely right-censored oropharyngeal head and neck cancer patients for discriminative groupings to improve outcome prediction*, Scientific reports, 10 (2020), pp. 1–14.

[269] J. TRELLES TRABUCCO, D. LEE, S. DERRIBLE, AND G. E. MARAI, *Visual analysis of a smart city‚Äôs energy consumption*, Multimodal Technologies and Interaction, 3 (2019), p. 30.

[270] A. ULBINAITE, M. KUCINSKIENE, AND Y. LE MOULLEC, *Determinants of insurance purchase decision making in lithuania*, Engineering Economics, 24 (2013), pp. 144–159.

[271] L. VAN DER MAATEN AND G. HINTON, *Visualizing data using t-sne.*, Journal of machine learning research, 9 (2008).

[272] D. VARGA, *Fintech, the new era of financial services*, Vezetéstudomány-Budapest Management Review, 48 (2017), pp. 22–32.

[273] A. VELLIDO, *The importance of interpretability and visualization in machine learning for applications in medicine and health care*, Neural computing and applications, (2019), pp. 1–15.

[274] A. VIJ, S. RYAN, S. SAMPSON, AND S. HARRIS, *Consumer preferences for on-demand transport in australia*, Transportation Research Part A: Policy and Practice, 132 (2020), pp. 823–839.

[275] P. VISWANATHAN, S. SRINIVASAN, AND N. HARIHARAN, *Predicting financial health of banks for investor guidance using machine learning algorithms*, Journal of Emerging Market Finance, (2020), p. 0972652720913478.

[276] N. N. VO, S. LIU, X. LI, AND G. XU, *Leveraging unstructured call log data for customer churn prediction*, Knowledge-Based Systems, 212 (2021), p. 106586.

[277] T. VON LANDESBERGER, S. BREMM, M. KIRSCHNER, S. WESARG, AND A. KUIJPER, *Visual analytics for model-based medical image segmentation: Opportunities and challenges*, Expert Systems with Applications, 40 (2013), pp. 4934–4943.

[278] M. WAGNER, F. FISCHER, R. LUH, A. HABERSON, A. RIND, D. A. KEIM, AND W. AIGNER, *A survey of visualization systems for malware analysis*, in Eurographics Conference on Visualization (EuroVis), 2015, pp. 105–125.

[279] J. WALKER, R. BORGO, AND M. W. JONES, *Timenotes: a study on effective chart visualization and interaction techniques for time-series data*, IEEE transactions on visualization and computer graphics, 22 (2015), pp. 549–558.

[280] J. WANG, *Interpreting and Diagnosing Deep Learning Models: A Visual Analytics Approach*, PhD thesis, The Ohio State University, 2019.

[281] J. WANG, L. GOU, H.-W. SHEN, AND H. YANG, *Dqnviz: A visual analytics approach to understand deep q-networks*, IEEE transactions on visualization and computer graphics, 25 (2018), pp. 288–298.

[282] J. WANG, L. GOU, W. ZHANG, H. YANG, AND H.-W. SHEN, *Deepvid: Deep visual interpretation and diagnosis for image classifiers via knowledge distillation*, IEEE transactions on visualization and computer graphics, 25 (2019), pp. 2168–2180.

[283] J. WANG, W. ZHANG, AND H. YANG, *Scanviz: Interpreting the symbol-concept association captured by deep neural networks through visual analytics*, in 2020 IEEE Pacific Visualization Symposium (PacificVis), IEEE, 2020, pp. 51–60.

[284] Y. WANG, L. ZHANG, M. NIU, R. LI, R. TU, X. LIU, J. HOU, Z. MAO, Z. WANG, AND C. WANG, *Genetic risk score increased discriminant efficiency of predictive models for type 2 diabetes mellitus using machine learning: Cohort study*, Frontiers in public health, 9 (2021), p. 96.

[285] Z. WANG, Q. LI, G. LI, AND G. XU, *Polynomial representation for persistence diagram*, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 6123–6132.

[286] Z. J. WANG, R. TURKO, AND D. H. CHAU, *Dodrio: Exploring transformer models with interactive visualization*, arXiv preprint arXiv:2103.14625, (2021).

[287] W. N. WASSOUF, R. ALKHATIB, K. SALLOUM, AND S. BALLOUL, *Predictive analytics using big data for increased customer loyalty: Syriatel telecom company case study*, Journal of Big Data, 7 (2020), pp. 1–24.

[288] W. WHITTAKER, M. SUTTON, S. MACDONALD, M. MAXWELL, M. SMITH, P. WILSON, AND J. MORRISON, *The effect of mental ill health on absence from work in different occupational classifications: analysis of routine data in the british household panel survey*, Journal of occupational and environmental medicine, 54 (2012), pp. 1539–1544.

[289] J. C. WONG, Z. IZADI, S. SCHROEDER, M. NADER, J. MIN, A. B. NEINSTEIN, AND S. ADI, *A pilot study of use of a software platform for the collection, integration, and visualization of diabetes device data by health care providers in a multidisciplinary pediatric setting*, Diabetes technology & therapeutics, 20 (2018), pp. 806–816.

[290] J. C. WONG, A. B. NEINSTEIN, H. LOOK, B. ARBITER, N. CHOKR, C. ROSS, AND S. ADI, *Pilot study of a novel application for data visualization in type 1 diabetes*, Journal of diabetes science and technology, 11 (2017), pp. 800–807.

[291] K. WONGSUPHASAWAT, D. SMILKOV, J. WEXLER, J. WILSON, D. MANE, D. FRITZ, D. KRISHNAN, F. B. VIÉGAS, AND M. WATTENBERG, *Visualizing dataflow graphs of deep learning models in tensorflow*, IEEE transactions on visualization and computer graphics, 24 (2017), pp. 1–12.

[292] S. WU AND S. WANG, *Information-theoretic outlier detection for large-scale categorical data*, IEEE transactions on knowledge and data engineering, 25 (2011), pp. 589–602.

[293] M. V. WUTHRICH AND C. BUSER, *Data analytics for non-life insurance pricing*, Swiss Finance Institute Research Paper, (2020).

[294] C. XIE, W. CHEN, X. HUANG, Y. HU, S. BARLOWE, AND J. YANG, *Vaet: A visual analytics approach for e-transactions time-series*, IEEE transactions on visualization and computer graphics, 20 (2014), pp. 1743–1752.

[295] K. YADA, H. MOTODA, T. WASHIO, AND A. MIYAWAKI, *Consumer behavior analysis by graph mining technique*, New Mathematics and Natural Computation, 2 (2006), pp. 59–68.

[296] A. YAELI, P. BAK, G. FEIGENBLAT, S. NADLER, H. ROITMAN, G. SAADOUN, H. J. SHIP, D. COHEN, O. FUCHS, S. OFEK-KOIFMAN, ET AL., *Understanding customer behavior using indoor location analysis and visualization*, IBM Journal of Research and Development, 58 (2014), pp. 3–1.

[297] G. YANG, *The complexity of mining maximal frequent itemsets and maximal frequent patterns*, in Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, 2004, pp. 344–353.

[298] G.-E. YAP, A.-H. TAN, AND H.-H. PANG, *Explaining inferences in bayesian networks*, Applied Intelligence, 29 (2008), pp. 263–278.

[299] R. YASMIN, M. SALMINEN, E. GILMAN, J. PETÄJÄJÄRVI, K. MIKHAYLOV, M. PAKANEN, A. NIEMELÄ, J. RIEKKI, S. PIRTTIKANGAS, AND A. POUTTU, *Combining iot deployment and data visualization: experiences within campus maintenance use-case*, in 2018 9th International Conference on the Network of the Future (NOF), IEEE, 2018, pp. 101–105.

[300] P. C. YAU, D. WONG, W. H. LUEN, AND J. LEUNG, *Understanding consumer behavior by big data visualization in the smart space laboratory*, in Proceedings

of the 2020 5th International Conference on Big Data and Computing, 2020, pp. 13–17.

[301] J. YIN, Q. LI, S. LIU, Z. WU, AND G. XU, *Leveraging multi-level dependency of relational sequences for social spammer detection*, arXiv preprint arXiv:2009.06231, (2020).

[302] J. YIN, Z. ZHOU, S. LIU, Z. WU, AND G. XU, *Social spammer detection: A multi-relational embedding approach*, in Pacific-Asia Conference on Knowledge Discovery and Data Mining, Springer, 2018, pp. 615–627.

[303] R. K. YIN, *Case study research: Design and methods*, vol. 5, sage, 2009.

[304] J. YOSINSKI, J. CLUNE, A. NGUYEN, T. FUCHS, AND H. LIPSON, *Understanding neural networks through deep visualization*, arXiv preprint arXiv:1506.06579, (2015).

[305] B. YU AND C. T. SILVA, *Flowsense: A natural language interface for visual data exploration within a dataflow system*, IEEE transactions on visualization and computer graphics, 26 (2019), pp. 1–11.

[306] C. YU, J. LI, X. LI, X. REN, AND B. B. GUPTA, *Four-image encryption scheme based on quaternion fresnel transform, chaos and computer generated hologram*, Multimedia Tools and Applications, 77 (2018), pp. 4585–4608.

[307] X. YU, S. GUO, J. GUO, AND X. HUANG, *An extended support vector machine forecasting framework for customer churn in e-commerce*, Expert Systems with Applications, 38 (2011), pp. 1425–1430.

[308] X. YUE, *Financial intelligence and strategy extraction via visual analysis approaches*, PhD thesis, 2019.

[309] X. YUE, X. SHU, X. ZHU, X. DU, Z. YU, D. PAPADOPOULOS, AND S. LIU, *Bitextract: Interactive visualization for extracting bitcoin exchange intelligence*, IEEE transactions on visualization and computer graphics, 25 (2018), pp. 162–171.

[310] P. YÜKSEL AND S. YILDIRIM, *Theoretical frameworks, methods, and procedures for conducting phenomenological studies in educational settings*, Turkish online journal of qualitative inquiry, 6 (2015), pp. 1–20.

[311] H. ZENG, *Towards better understanding of deep learning with visualization*, The Hong Kong University of Science and Technology, (2016).

[312] W. ZENG, C. LIN, J. LIN, J. JIANG, J. XIA, C. TURKAY, AND W. CHEN, *Revisiting the modifiable areal unit problem in deep traffic prediction with visual analytics*, IEEE Transactions on Visualization and Computer Graphics, (2020).

[313] J. ZERAFA, M. R. ISLAM, A. KABIR, AND G. XU, *Extravis: Exploration of traffic incidents using visual interactive system*, in 25th International Conference Information Visualisation (IV 2021), IEEE, Institute of Electrical and Electronics Engineers, 2021.

[314] K. ZHANG, Y. CHEN, AND C. LI, *Discovering the tourists' behaviors and perceptions in a tourism destination by analyzing photos' visual content with a computer deep learning model: The case of beijing*, Tourism Management, 75 (2019), pp. 595–608.

[315] S. ZHANG AND D. METAXAS, *Large-scale medical image analytics: Recent methodologies, applications and future directions*, 2016.

[316] Y. ZHANG, K. CHANANA, AND C. DUNNE, *Idmvis: Temporal event sequence visualization for type 1 diabetes treatment decision support*, IEEE transactions on visualization and computer graphics, 25 (2018), pp. 512–522.

[317] Y. ZHANG AND X. CHEN, *Explainable recommendation: A survey and new perspectives*, arXiv preprint arXiv:1804.11192, (2018).

[318] Z. ZHOU, S. LIU, G. XU, AND W. ZHANG, *On completing sparse knowledge base with transitive relation embedding*, in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, 2019, pp. 3125–3132.

[319] P. ZIMMET, K. G. ALBERTI, D. J. MAGLIANO, AND P. H. BENNETT, *Diabetes mellitus statistics on prevalence and mortality: facts and fallacies*, Nature Reviews Endocrinology, 12 (2016), pp. 616–622.

[320] L. M. ZINTGRAF, T. S. COHEN, T. ADEL, AND M. WELLING, *Visualizing deep neural network decisions: Prediction difference analysis*, arXiv preprint arXiv:1702.04595, (2017).